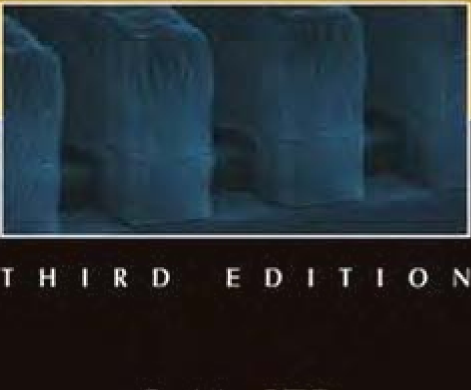
Copyrighted Material

WILEY

PHYSICS OF SEMICONDUCTOR DEVICES



THIRD EDITION

S. M. SZE
KWOK K. NG

Copyrighted Material

VISIT…

LANZAROTE
Caliente.COM

Physics of Semiconductor Devices

Third Edition

S. M. Sze
Department of Electronics Engineering
National Chiao Tung University
Hsinchu, Taiwan

and

Kwok K. Ng
Central Laboratory
MVC (a subsidiary of ProMOS Technologies, Taiwan)
San Jose, California

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Description of cover photograph:

A scanning electron micrograph of an array of the floating-gate nonvolatile semiconductor memory (NVSM) magnified 100,000 times. NVSM was invented at Bell Telephone Laboratories in 1967. There are more NVSM cells produced annually in the world than any other semiconductor device and, for that matter, any other human-made item. For a discussion of this device, see Chapter 6. Photo courtesy of Macronix International Company, Hsinchu, Taiwan, ROC.

Copyright © 2007 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data is available.

ISBN-13: 978-0-471-14323-9
ISBN-10: 0-471-14323-5

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Preface

Since the mid-20th Century the electronics industry has enjoyed phenomenal growth and is now the largest industry in the world. The foundation of the electronics industry is the *semiconductor device*. To meet the tremendous demand of this industry, the semiconductor-device field has also grown rapidly. Coincident with this growth, the semiconductor-device literature has expanded and diversified. For access to this massive amount of information, there is a need for a book giving a comprehensive introductory account of device physics and operational principles.

With the intention of meeting such a need, the First Edition and the Second Edition of *Physics of Semiconductor Devices* were published in 1969 and 1981, respectively. It is perhaps somewhat surprising that the book has so long held its place as one of the main textbooks for advanced undergraduate and graduate students in applied physics, electrical and electronics engineering, and materials science. Because the book includes much useful information on material parameters and device physics, it is also a major reference for engineers and scientists in semiconductor-device research and development. To date, the book is one of the most, if not *the* most, cited works in contemporary engineering and applied science with over 15,000 citations (ISI, Thomson Scientific).

Since 1981, more than 250,000 papers on semiconductor devices have been published, with numerous breakthroughs in device concepts and performances. The book clearly needed another major revision if it were to continue to serve its purpose. In this Third Edition of *Physics of Semiconductor Devices*, over 50% of the material has been revised or updated, and the material has been totally reorganized. We have retained the basic physics of classic devices and added many sections that are of contemporary interest such as the three-dimensional MOSFETs, nonvolatile memory, modulation-doped field-effect transistor, single-electron transistor, resonant-tunneling diode, insulated-gate bipolar transistor, quantum cascade laser, semiconductor sensors, and so on. On the other hand, we have omitted or reduced sections of less-important topics to maintain the overall book length.

We have added a problem set at the end of each chapter. The problem set forms an integral part of the development of the topics, and some problems can be used as worked examples in the classroom. A complete set of detailed solutions to all end-of-chapter problems has been prepared. The solution manuals are available free to all adopting faculties. The figures and tables used in the text are also available, in electronic format, to instructors from the publisher. Instructors can find out more information at the publisher's website at *http://www.wiley.com/interscience/sze.*

In the course of writing this text, we had the fortune of help and support of many people. First we express our gratitude to the management of our academic and industrial institutions, the National Chiao Tung University, the National Nano Device Laboratories, Agere Systems, and MVC, without whose support this book could not have been written. We wish to thank the Spring Foundation of the National Chiao Tung University for the financial support. One of us (K. Ng) would like to thank J. Hwang and B. Leung for their continued encouragement and personal help.

We have benefited greatly from suggestions made by our reviewers who took their time from their busy schedule. Credits are due to the following scholars: A. Alam, W. Anderson, S. Banerjee, J. Brews, H. C. Casey, Jr., P. Chow, N. de Rooij, H. Eisele, E. Kasper, S. Luryi, D. Monroe, P. Panayotatos, S. Pearton, E. F. Schubert, A. Seabaugh, M. Shur, Y. Taur, M. Teich, Y. Tsividis, R. Tung, E. Yang, and A. Zaslavsky. We also appreciate the permission granted to us from the respective journals and authors to reproduce their original figures cited in this work.

It is our pleasure to acknowledge the help of many family members in preparing the manuscript in electronic format; Kyle Eng and Valerie Eng in scanning and importing text from the Second Edition, Vivian Eng in typing the equations, and Jennifer Tao in preparing the figures which have all been redrawn. We are further thankful to Norman Erdos for technical editing of the entire manuscript, and to Iris Lin and Nai-Hua Chang for preparing the problem sets and solution manual. At John Wiley and Sons, we wish to thank George Telecki who encouraged us to undertake the project. Finally, we are grateful to our wives, Therese Sze and Linda Ng, for their support and assistance during the course of the book project.

S. M. Sze
Hsinchu, Taiwan

Kwok K. Ng
San Jose, California
July 2006

Contents

Introduction 1

Part I Semiconductor Physics

Chapter 1 Physics and Properties of Semiconductors—A Review 7

1.1 Introduction, 7
1.2 Crystal Structure, 8
1.3 Energy Bands and Energy Gap, 12
1.4 Carrier Concentration at Thermal Equilibrium, 16
1.5 Carrier-Transport Phenomena, 28
1.6 Phonon, Optical, and Thermal Properties, 50
1.7 Heterojunctions and Nanostructures, 56
1.8 Basic Equations and Examples, 62

Part II Device Building Blocks

Chapter 2 *p-n* Junctions 79

2.1 Introduction, 79
2.2 Depletion Region, 80
2.3 Current-Voltage Characteristics, 90
2.4 Junction Breakdown, 102
2.5 Transient Behavior and Noise, 114
2.6 Terminal Functions, 118
2.7 Heterojunctions, 124

Chapter 3 Metal-Semiconductor Contacts 134

3.1 Introduction, 134
3.2 Formation of Barrier, 135
3.3 Current Transport Processes, 153
3.4 Measurement of Barrier Height, 170
3.5 Device Structures, 181
3.6 Ohmic Contact, 187

Chapter 4 Metal-Insulator-Semiconductor Capacitors **197**

4.1 Introduction, 197
4.2 Ideal MIS Capacitor, 198
4.3 Silicon MOS Capacitor, 213

Part III Transistors

Chapter 5 Bipolar Transistors **243**

5.1 Introduction, 243
5.2 Static Characteristics, 244
5.3 Microwave Characteristics, 262
5.4 Related Device Structures, 275
5.5 Heterojunction Bipolar Transistor, 282

Chapter 6 MOSFETs **293**

6.1 Introduction, 293
6.2 Basic Device Characteristics, 297
6.3 Nonuniform Doping and Buried-Channel Device, 320
6.4 Device Scaling and Short-Channel Effects, 328
6.5 MOSFET Structures, 339
6.6 Circuit Applications, 347
6.7 Nonvolatile Memory Devices, 350
6.8 Single-Electron Transistor, 360

Chapter 7 JFETs, MESFETs, and MODFETs **374**

7.1 Introduction, 374
7.2 JFET and MESFET, 375
7.3 MODFET, 401

Part IV Negative-Resistance and Power Devices

Chapter 8 Tunnel Devices **417**

8.1 Introduction, 417
8.2 Tunnel Diode, 418
8.3 Related Tunnel Devices, 435
8.4 Resonant-Tunneling Diode, 454

Chapter 9 IMPATT Diodes **466**

9.1 Introduction, 466

9.2 Static Characteristics, 467
9.3 Dynamic Characteristics, 474
9.4 Power and Efficiency, 482
9.5 Noise Behavior, 489
9.6 Device Design and Performance, 493
9.7 BARITT Diode, 497
9.8 TUNNETT Diode, 504

Chapter 10 Transferred-Electron and Real-Space-Transfer Devices 510

10.1 Introduction, 510
10.2 Transferred-Electron Device, 511
10.3 Real-Space-Transfer Devices, 536

Chapter 11 Thyristors and Power Devices 548

11.1 Introduction, 548
11.2 Thyristor Characteristics, 549
11.3 Thyristor Variations, 574
11.4 Other Power Devices, 582

Part V Photonic Devices and Sensors

Chapter 12 LEDs and Lasers 601

12.1 Introduction, 601
12.2 Radiative Transitions, 603
12.3 Light-Emitting Diode (LED), 608
12.4 Laser Physics, 621
12.5 Laser Operating Characteristics, 630
12.6 Specialty Lasers, 651

Chapter 13 Photodetectors and Solar Cells 663

13.1 Introduction, 663
13.2 Photoconductor, 667
13.3 Photodiodes, 671
13.4 Avalanche Photodiode, 683
13.5 Phototransistor, 694
13.6 Charge-Coupled Device (CCD), 697
13.7 Metal-Semiconductor-Metal Photodetector, 712
13.8 Quantum-Well Infrared Photodetector, 716
13.9 Solar Cell, 719

Chapter 14 Sensors 743

14.1 Introduction, 743
14.2 Thermal Sensors, 744
14.3 Mechanical Sensors, 750
14.4 Magnetic Sensors, 758
14.5 Chemical Sensors, 765

Appendixes 773

A. List of Symbols, 775
B. International System of Units, 785
C. Unit Prefixes, 786
D. Greek Alphabet, 787
E. Physical Constants, 788
F. Properties of Important Semiconductors, 789
G. Properties of Si and GaAs, 790
H. Properties of SiO_2 and Si_3N_4, 791

Index 793

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

Introduction

The book is organized into five parts:

Part I: Semiconductor Physics
Part II: Device Building Blocks
Part III: Transistors
Part IV: Negative-Resistance and Power Devices
Part V: Photonic Devices and Sensors

Part I, Chapter 1, is a summary of semiconductor properties that are used throughout the book as a basis for understanding and calculating device characteristics. Energy band, carrier concentration, and transport properties are briefly surveyed, with emphasis on the two most-important semiconductors: silicon (Si) and gallium arsenide (GaAs). A compilation of the recommended or most-accurate values for these semiconductors is given in the illustrations of Chapter 1 and in the Appendixes for convenient reference.

Part II, Chapters 2 through 4, treats the basic device building blocks from which all semiconductor devices can be constructed. Chapter 2 considers the *p-n* junction characteristics. Because the *p-n* junction is the building block of most semiconductor devices, *p-n* junction theory serves as the foundation of the physics of semiconductor devices. Chapter 2 also considers the heterojunction, that is a junction formed between two dissimilar semiconductors. For example, we can use gallium arsenide (GaAs) and aluminum arsenide (AlAs) to form a heterojunction. The heterojunction is a key building block for high-speed and photonic devices. Chapter 3 treats the metal-semiconductor contact, which is an intimate contact between a metal and a semiconductor. The contact can be rectifying similar to a *p-n* junction if the semiconductor is moderately doped and becomes ohmic if the semiconductor is very heavily doped. An ohmic contact can pass current in either direction with a negligible voltage drop and can provide the necessary connections between devices and the outside world. Chapter 4 considers the metal-insulator-semiconductor (MIS) capacitor of which the Si-based metal-oxide-semiconductor (MOS) structure is the dominant member. Knowledge of the surface physics associated with the MOS capacitor is important, not only for understanding MOS-related devices such as the MOSFET and the floating-gate nonvolatile memory but also because of its relevance to the stability and reliability of all other semiconductor devices in their surface and isolation areas.

Part III, Chapters 5 through 7, deals with the transistor family. Chapter 5 treats the bipolar transistor, that is, the interaction between two closely coupled *p-n* junctions. The bipolar transistor is one of the most-important original semiconductor devices. The invention of the bipolar transistor in 1947 ushered in the modern electronic era. Chapter 6 considers the MOSFET (MOS field-effect transistor). The distinction between a field-effect transistor and a potential-effect transistor (such as the bipolar transistor) is that in the former, the channel is modulated by the gate through a capacitor whereas in the latter, the channel is controlled by a direct contact to the channel region.[1] The MOSFET is the most-important device for advanced integrated circuits, and is used extensively in microprocessors and DRAMs (dynamic random access memories). Chapter 6 also treats the nonvolatile semiconductor memory which is the dominant memory for portable electronic systems such as the cellular phone, notebook computer, digital camera, audio and video players, and global positioning system (GPS). Chapter 7 considers three other field-effect transistors; the JFET (junction field-effect-transistor), MESFET (metal-semiconductor field-effect transistor), and MODFET (modulation-doped field-effect transistor). The JFET is an older member and now used mainly as power devices, whereas the MESFET and MODFET are used in high-speed, high-input-impedance amplifiers and monolithic microwave integrated circuits.

Part IV, Chapters 8 through 11, considers negative-resistance and power devices. In Chapter 8, we discuss the tunnel diode (a heavily doped *p-n* junction) and the resonant-tunneling diode (a double-barrier structure formed by multiple heterojunctions). These devices show negative differential resistances due to quantum-mechanical tunneling. They can generate microwaves or serve as functional devices, that is, they can perform a given circuit function with a greatly reduced number of components. Chapter 9 discusses the transit-time devices. When a *p-n* junction or a metal-semiconductor junction is operated in avalanche breakdown, under proper conditions we have an IMPATT diode that can generate the highest CW (continuous wave) power output of all solid-state devices at millimeter-wave frequencies (i.e., above 30 GHz). The operational characteristics of the related BARITT and TUNNETT diodes are also presented. The transferred-electron device (TED) is considered in Chapter 10. Microwave oscillation can be generated by the mechanism of electron transfer from a high-mobility lower-energy valley in the conduction band to a low-mobility higher-energy valley (in momentum space), the transferred-electron effect. Also presented are the real-space-transfer devices which are similar to TED but the electron transfer occurs between a narrow-bandgap material to an adjacent wide-bandgap material in real space as opposed to momentum space. The thyristor, which is basically three closely coupled *p-n* junctions in the form of a *p-n-p-n* structure, is discussed in Chapter 11. Also considered are the MOS-controlled thyristor (a combination of MOSFET with a conventional thyristor) and the insulated-gate bipolar transistor (IGBT, a combination of MOSFET with a conventional bipolar transistor). These devices have a wide range of power-handling and switching capability; they can handle currents from a few milliamperes to thousands of amperes and voltages above 5000 V.

Part V, Chapters 12 through 14, treats photonic devices and sensors. Photonic devices can detect, generate, and convert optical energy to electric energy, or vice versa. The semiconductor light sources—light-emitting diode (LED) and laser, are discussed in Chapter 12. The LEDs have a multitude of applications as display devices such as in electronic equipment and traffic lights, and as illuminating devices such as flashlights and automobile headlights. Semiconductor lasers are used in optical-fiber communication, video players, and high-speed laser printing. Various photodetectors with high quantum efficiency and high response speed are discussed in Chapter 13. The chapter also considers the solar cell which converts optical energy to electrical energy similar to a photodetector but with different emphasis and device configuration. As the worldwide energy demand increases and the fossil-fuel supply will be exhausted soon, there is an urgent need to develop alternative energy sources. The solar cell is considered a major candidate because it can convert sunlight directly to electricity with good conversion efficiency, can provide practically everlasting power at low operating cost, and is virtually nonpolluting. Chapter 14 considers important semiconductor sensors. A sensor is defined as a device that can detect or measure an external signal. There are basically six types of signals: electrical, optical, thermal, mechanical, magnetic, and chemical. The sensors can provide us with informations about these signals which could not otherwise be directly perceived by our senses. Based on the definition of sensors, all traditional semiconductor devices are sensors since they have inputs and outputs and both are in electrical forms. We have considered the sensors for electrical signals in Chapters 2 through 11, and the sensors for optical signals in Chapters 12 and 13. In Chapter 14, we are concerned with sensors for the remaining four types of signals, i.e., thermal, mechanical, magnetic, and chemical.

We recommend that readers first study semiconductor physics (Part I) and the device building blocks (Part II) before moving to subsequent parts of the book. Each chapter in Parts III through V deals with a major device or a related device family, and is more or less independent of the other chapters. So, readers can use the book as a reference and instructors can select chapters appropriate for their classes and in their order of preference. We have a vast literature on semiconductor devices. To date, more than 300,000 papers have been published in this field, and the grand total may reach one million in the next decade. In this book, each chapter is presented in a clear and coherent fashion without heavy reliance on the original literature. However, we have an extensive listing of key papers at the end of each chapter for reference and for further reading.

REFERENCE

1. K. K. Ng, *Complete Guide to Semiconductor Devices*, 2nd Ed., Wiley, New York, 2002.

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

PART I

SEMICONDUCTOR PHYSICS

◆ **Chapter 1 Physics and Properties of Semiconductors —A Review**

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

1

Physics and Properties of Semiconductors—A Review

1.1 INTRODUCTION
1.2 CRYSTAL STRUCTURE
1.3 ENERGY BANDS AND ENERGY GAP
1.4 CARRIER CONCENTRATION AT THERMAL EQUILIBRIUM
1.5 CARRIER-TRANSPORT PHENOMENA
1.6 PHONON, OPTICAL, AND THERMAL PROPERTIES
1.7 HETEROJUNCTIONS AND NANOSTRUCTURES
1.8 BASIC EQUATIONS AND EXAMPLES

1.1 INTRODUCTION

The physics of semiconductor devices is naturally dependent on the physics of semiconductor materials themselves. This chapter presents a summary and review of the basic physics and properties of semiconductors. It represents only a small cross section of the vast literature on semiconductors; only those subjects pertinent to device operations are included here. For detailed consideration of semiconductor physics, the reader should consult the standard textbooks or reference works by Dunlap,[1] Madelung,[2] Moll,[3] Moss,[4] Smith.[5] Böer,[6] Seeger,[7] and Wang,[8] to name a few.

To condense a large amount of information into a single chapter, four tables (some in appendixes) and over 30 illustrations drawn from experimental data are compiled and presented here. This chapter emphasizes the two most-important semiconductors: silicon (Si) and gallium arsenide (GaAs). Silicon has been studied extensively and widely used in commercial electronics products. Gallium arsenide has been intensively investigated in recent years. Particular properties studied are its direct bandgap for photonic applications and its intervalley-carrier transport and higher mobility for generating microwaves.

1.2 CRYSTAL STRUCTURE

1.2.1 Primitive Cell and Crystal Plane

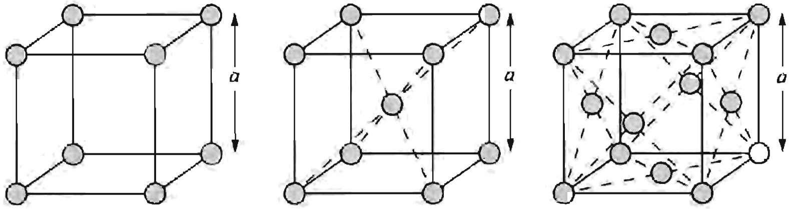
A crystal is characterized by having a well-structured periodic placement of atoms. The smallest assembly of atoms that can be repeated to form the entire crystal is called a primitive cell, with a dimension of lattice constant a. Figure 1 shows some important primitive cells.

Many important semiconductors have diamond or zincblende lattice structures which belong to the tetrahedral phases; that is, each atom is surrounded by four equidistant nearest neighbors which lie at the corners of a tetrahedron. The bond between two nearest neighbors is formed by two electrons with opposite spins. The diamond and the zincblende lattices can be considered as two interpenetrating face-centered cubic (fcc) lattices. For the diamond lattice, such as silicon (Fig. 1d), all the atoms are the same; whereas in a zincblende lattice, such as gallium arsenide (Fig. 1e), one sublattice is gallium and the other is arsenic. Gallium arsenide is a III-V compound, since it is formed from elements of groups III and V of the periodic table.

Most III-V compounds crystallize in the zincblende structure;[2,9] however, many semiconductors (including some III-V compounds) crystallize in the rock-salt or wurtzite structures. Figure 1f shows the rock-salt lattice, which again can be considered as two interpenetrating face-centered cubic lattices. In this rock-salt structure, each atom has six nearest neighbors. Figure 1g shows the wurtzite lattice, which can be considered as two interpenetrating hexagonal close-packed lattices (e.g., the sublattices of cadmium and sulfur). In this picture, for each sublattice (Cd or S), the two planes of adjacent layers are displaced horizontally such that the distance between these two planes are at a minimum (for a fixed distance between centers of two atoms), hence the name *close-packed*. The wurtzite structure has a tetrahedral arrangement of four equidistant nearest neighbors, similar to a zincblende structure.

Appendix F gives a summary of the lattice constants of important semiconductors, together with their crystal structures.[10,11] Note that some compounds, such as zinc sulfide and cadmium sulfide, can crystallize in either zincblende or wurtzite structures.

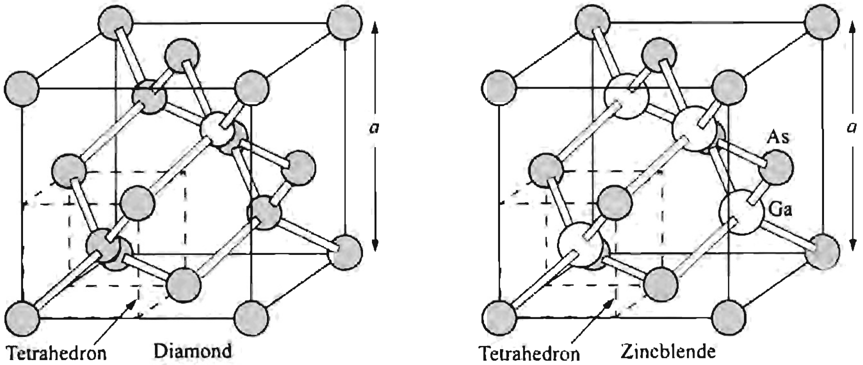
Since semiconductor devices are built on or near the semiconductor surface, the orientations and properties of the surface crystal planes are important. A convenient method of defining the various planes in a crystal is to use Miller indices. These indices are determined by first finding the intercepts of the plane with the three basis axes in terms of the lattice constants (or primitive cells), and then taking the reciprocals of these numbers and reducing them to the smallest three integers having the same ratio. The result is enclosed in parentheses (hkl) called the Miller indices for a single plane or a set of parallel planes $\{hkl\}$. Figure 2 shows the Miller indices of important planes in a cubic crystal. Some other conventions are given in Table 1. For



Simple cubic (Po) **(a)**

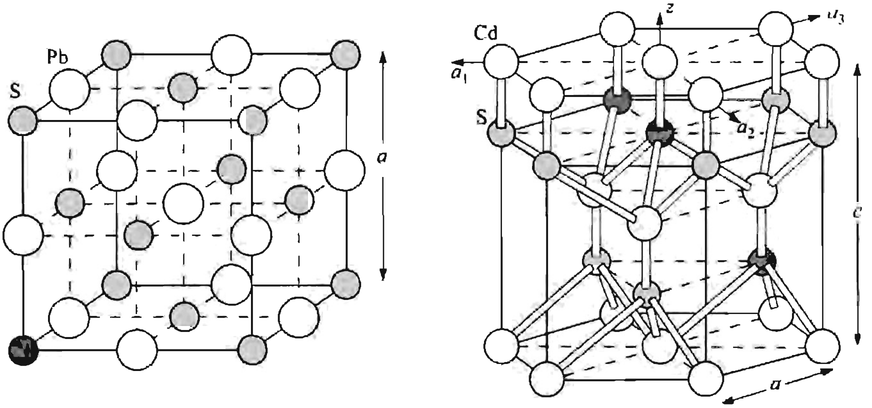
Body-centered cubic (Na, W, etc.) **(b)**

Face-centered cubic (Al, Au, etc.) **(c)**


a
As
Ga
a
Tetrahedron
Tetrahedron

Diamond (Si, Ge, C, etc.) **(d)**

Zincblende (GaAs, GaP, etc.) **(e)**


Pb
S
a
Cd
z
a_3
a_1
S
a_2
c
a

Rock-salt (PbS, PbTe, etc.) **(f)**

Wurtzite (CdS, ZnS, etc.) **(g)**

Fig. 1 Some important primitive cells (direct lattices) and their representative elements; a is the lattice constant.

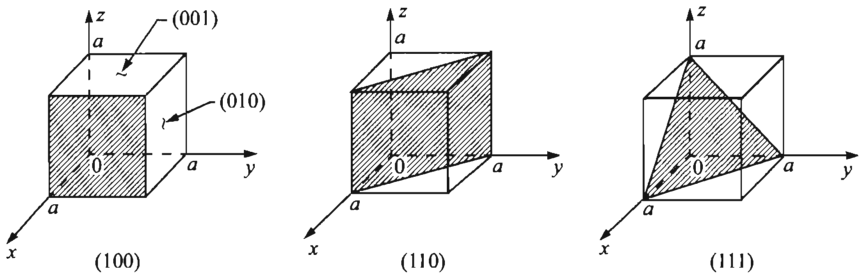


Fig. 2 Miller indices of some important planes in a cubic crystal.

silicon, a single-element semiconductor, the easiest breakage or cleavage planes are the {111} planes. In contrast, gallium arsenide, which has a similar lattice structure but also has a slight ionic component in the bonds, cleaves on {110} planes.

Three primitive basis vectors, ***a***, ***b***, and ***c*** of a primitive cell, describe a crystalline solid such that the crystal structure remains invariant under translation through any vector that is the sum of integral multiples of these basis vectors. In other words, the direct lattice sites can be defined by the set[12]

$$\boldsymbol{R} = m\boldsymbol{a} + n\boldsymbol{b} + p\boldsymbol{c} \tag{1}$$

where m, n, and p are integers.

1.2.2 Reciprocal Lattice

For a given set of the direct basis vectors, a set of reciprocal lattice basis vectors $\boldsymbol{a}^*$, $\boldsymbol{b}^*$, $\boldsymbol{c}^*$ can be defined as

$$\boldsymbol{a}^* \equiv 2\pi \frac{\boldsymbol{b} \times \boldsymbol{c}}{\boldsymbol{a} \cdot \boldsymbol{b} \times \boldsymbol{c}}, \tag{2a}$$

$$\boldsymbol{b}^* \equiv 2\pi \frac{\boldsymbol{c} \times \boldsymbol{a}}{\boldsymbol{a} \cdot \boldsymbol{b} \times \boldsymbol{c}}, \tag{2b}$$

Table 1 Miller Indices and Their Represented Plane or Direction of a Crystal Surface

Miller Indices	Description of plane or direction
(hkl)	For a plane that intercepts $1/h$, $1/k$, $1/l$ on the x-, y-, and z-axis, respectively.
$(\bar{h}kl)$	For a plane that intercepts the negative x-axis.
$\{hkl\}$	For a full set of planes of equivalent symmetry, such as {100} for (100), (010), (001), $(\bar{1}00)$, $(0\bar{1}0)$, and $(00\bar{1})$ in cubic symmetry.
$[hkl]$	For a direction of a crystal such as [100] for the x-axis.
$\langle hkl\rangle$	For a full set of equivalent directions.
$[hklm]$	For a plane in a hexagonal lattice (such as wurtzite) that intercepts $1/h$, $1/k$, $1/l$, $1/m$ on the a_1-, a_2-, a_3-, and z-axis, respectively (Fig. 1g).

$$c^* \equiv 2\pi \frac{\boldsymbol{a} \times \boldsymbol{b}}{\boldsymbol{a} \cdot \boldsymbol{b} \times \boldsymbol{c}} \tag{2c}$$

such that $\boldsymbol{a} \cdot \boldsymbol{a}^* = 2\pi$, $\boldsymbol{a} \cdot \boldsymbol{b}^* = 0$, and so on. The denominators are identical due to the equality that $\boldsymbol{a} \cdot \boldsymbol{b} \times \boldsymbol{c} = \boldsymbol{b} \cdot \boldsymbol{c} \times \boldsymbol{a} = \boldsymbol{c} \cdot \boldsymbol{a} \times \boldsymbol{b}$ which is the volume enclosed by these vectors. The general reciprocal lattice vector is given by

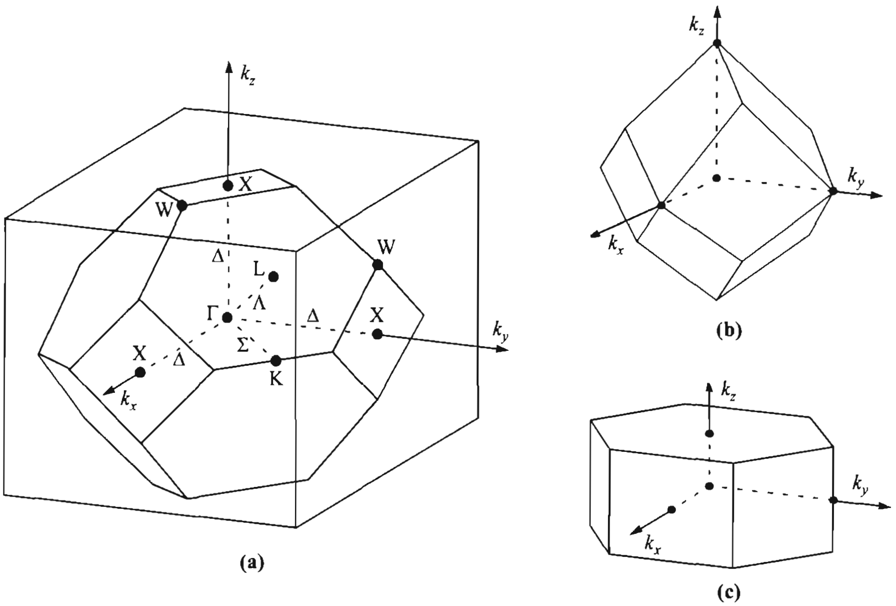
$$\boldsymbol{G} = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^* \tag{3}$$

where h, k, and l are integers. It follows that one important relationship between the direct lattice and the reciprocal lattice is

$$\boldsymbol{G} \cdot \boldsymbol{R} = 2\pi \times \text{Integer}, \tag{4}$$

and therefore each vector of the reciprocal lattice is normal to a set of planes in the direct lattice. The volume V_c^* of a primitive cell of the reciprocal lattice is inversely proportional to that (V_c) of the direct lattice; that is, $V_c^* = (2\pi)^3/V_c$, where $V_c \equiv \boldsymbol{a} \cdot \boldsymbol{b} \times \boldsymbol{c}$.

The primitive cell of a reciprocal lattice can be represented by a Wigner-Seitz cell. The Wigner-Seitz cell is constructed by drawing perpendicular bisector planes in the reciprocal lattice from the chosen center to the nearest equivalent reciprocal lattice sites. This technique can also be applied to a direct lattice. The Wigner-Seitz cell in the reciprocal lattice is called the first Brillouin zone. Figure 3a shows a typical example for a body-centered cubic (bcc) reciprocal lattice.[13] If one first draws lines from the center point (Γ) to the eight corners of the cube, then forms the bisector



(a)

(b)

(c)

Fig. 3 Brillouin zones for (a) fcc, diamond, and zincblende lattices, (b) bcc lattice, and (c) wurtzite lattice.

planes, the result is the truncated octahedron within the cube—a Wigner-Seitz cell. It can be shown that[14] a face-centered cubic (fcc) direct lattice with lattice constant a has a bcc reciprocal lattice with spacing $4\pi/a$. Thus the Wigner-Seitz cell shown in Fig. 3a is the primitive cell of the reciprocal (bcc) lattice for an fcc direct lattice. The Wigner-Seitz cells for bcc and hexagonal direct lattices can be similarly constructed and shown in Figs. 3b and 3c.[15] It will be shown that the reciprocal lattice is useful to visualize the E-k relationship when the coordinates of the wave vectors $\boldsymbol{k}$ ($|\boldsymbol{k}| = k = 2\pi/\lambda$) are mapped into the coordinates of the reciprocal lattice. In particular, the Brillouin zone for the fcc lattice is important because it is relevant to most semiconductor materials of interest here. The symbols used in Fig. 3a will be discussed in more details.

1.3 ENERGY BANDS AND ENERGY GAP

The energy-momentum (E-k) relationship for carriers in a lattice is important, for example, in the interactions with photons and phonons where energy and momentum have to be conserved, and with each other (electrons and holes) which leads to the concept of energy gap. This relationship also characterizes the effective mass and the group velocity, as will be discussed later.

The band structure of a crystalline solid, that is, the energy-momentum (E-k) relationship, is usually obtained by solving the Schrödinger equation of an approximate one-electron problem. The Bloch theorem, one of the most-important theorems basic to band structure, states that if a potential energy $V(\boldsymbol{r})$ is periodic in the direct lattice space, then the solutions for the wavefunction $\psi(\boldsymbol{r},\boldsymbol{k})$ of the Schrödinger equation[14,16]

$$\left[-\frac{\hbar^2}{2m^*}\nabla^2 + V(\boldsymbol{r})\right]\psi(\boldsymbol{r},\boldsymbol{k}) = E(\boldsymbol{k})\psi(\boldsymbol{r},\boldsymbol{k}) \tag{5}$$

are of the form of a Bloch function

$$\psi(\boldsymbol{r},\boldsymbol{k}) = \exp(j\boldsymbol{k}\cdot\boldsymbol{r})U_b(\boldsymbol{r},\boldsymbol{k}). \tag{6}$$

Here b is the band index, $\psi(\boldsymbol{r},\boldsymbol{k})$ and $U_b(\boldsymbol{r},\boldsymbol{k})$ are periodic in $\boldsymbol{R}$ of the direct lattice. Since

$$\begin{aligned}\psi(\boldsymbol{r}+\boldsymbol{R},\boldsymbol{k}) &= \exp[j\boldsymbol{k}\cdot(\boldsymbol{r}+\boldsymbol{R})]U_b(\boldsymbol{r}+\boldsymbol{R},\boldsymbol{k}) \\ &= \exp(j\boldsymbol{k}\cdot\boldsymbol{r})\exp(j\boldsymbol{k}\cdot\boldsymbol{R})U_b(\boldsymbol{r},\boldsymbol{k}),\end{aligned} \tag{7}$$

and is equal to $\psi(\boldsymbol{r},\boldsymbol{k})$, it is necessary that $\boldsymbol{k}\cdot\boldsymbol{R}$ is a multiple of 2π. It is the property of Eq. 4 that the reciprocal lattice can be used when $\boldsymbol{G}$ is replaced with $\boldsymbol{k}$ for visualizing the E-k relationship.

From the Bloch theorem one can also show that the energy $E(\boldsymbol{k})$ is periodic in the reciprocal lattice, that is, $E(\boldsymbol{k}) = E(\boldsymbol{k}+\boldsymbol{G})$, where $\boldsymbol{G}$ is given by Eq. 3. For a given band index, to label the energy uniquely, it is sufficient to use only $\boldsymbol{k}$'s in a primitive cell of the reciprocal lattice. The standard convention is to use the Wigner-Seitz cell in the reciprocal lattice (Fig. 3). This cell is the Brillouin zone or the first Brillouin zone.[13] It is thus evident that we can reduce any momentum $\boldsymbol{k}$ in the reciprocal space to a point inside the Brillouin zone, where any energy state can be given a label in the reduced zone schemes.

The Brillouin zone for the diamond and the zincblende lattices is the same as that of the fcc and is shown in Fig. 3a. Table 2 summarizes its most-important symmetry points and symmetry lines, such as the center of the zone, the zone edges and their corresponding k axes.

The energy bands of solids have been studied theoretically using a variety of numerical methods. For semiconductors the three methods most frequently used are the orthogonalized plane-wave method,[17,18] the pseudopotential method,[19] and the $\boldsymbol{k} \cdot \boldsymbol{p}$ method.[5] Figure 4 shows results of studies of the energy-band structures of Si and GaAs. Notice that for any semiconductor there is a forbidden energy range in which allowed states cannot exist. Energy regions or energy bands are permitted above and below this energy gap. The upper bands are called the conduction bands; the lower bands, the valence bands. The separation between the energy of the lowest conduction band and that of the highest valence band is called the bandgap or energy gap E_g, which is one of the most-important parameters in semiconductor physics. In this figure the bottom of the conduction band is designated E_C, and the top of the valence band E_V. Within the bands, the electron energy is conventionally defined to be positive when measured upward from E_C, and the hole energy is positive when measured downward from E_V. The bandgaps of some important semiconductors are listed in Appendix F.

The valence band in the zincblende structure, such as that for GaAs in Fig. 4b, consists of four subbands when spin is neglected in the Schrödinger equation, and each band is doubled when spin is taken into account. Three of the four bands are degenerate at $k = 0$ (Γ point) and form the upper edge of the band, and the fourth band forms the bottom (not shown). Furthermore, the spin-orbit interaction causes a splitting of the band at $k = 0$.

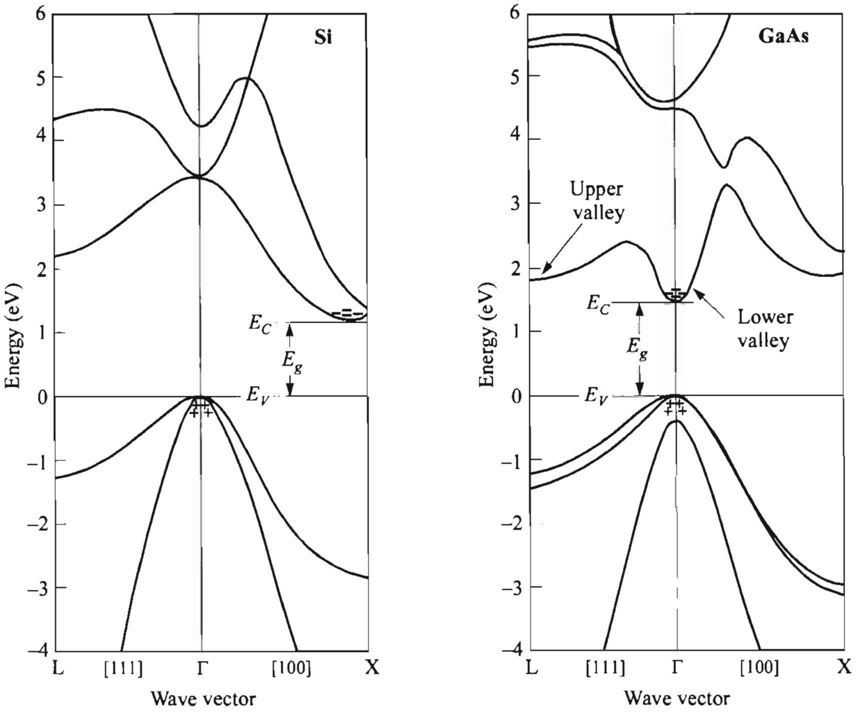
Near the band edges, i.e., bottom of E_C and top of E_V, the E-k relationship can be approximated by a quadratic equation

$$E(k) = \frac{\hbar^2 k^2}{2m^*}, \tag{8}$$

where m^* is the associated effective mass. But as shown in Fig. 4, along a given direction, the two top valence bands can be approximated by two parabolic bands with different curvatures: the heavy-hole band (the wider band in k-axis with smaller $\partial^2 E/\partial k^2$)

Table 2 Brillouin Zone of fcc, Diamond, and Zincblende Lattices: Zone Edges and Their Corresponding Axes (Γ is the Center)

Point	Degeneracy	Axis
Γ, (0,0,0)	1	
X, $2\pi/a(\pm1,0,0)$, $2\pi/a(0,\pm1,0)$, $2\pi/a(0,0,\pm1)$	6	Δ, ⟨1,0,0⟩
L, $2\pi/a(\pm1/2,\pm1/2,\pm1/2)$	8	Λ, ⟨1,1,1⟩
K, $2\pi/a(\pm3/4,\pm3/4,0)$, $2\pi/a(0,\pm3/4,\pm3/4)$, $2\pi/a(\pm3/4,0,\pm3/4)$	12	Σ, ⟨1,1,0⟩



(a) (b)

Fig. 4 Energy-band structures of (a) Si and (b) GaAs, where E_g is the energy bandgap. Plus signs (+) indicate holes in the valence bands and minus signs (–) indicate electrons in the conduction bands. (After Ref. 20.)

and the light-hole band (the narrower band with larger $\partial^2E/\partial k^2$). The effective mass in general is tensorial with components m_{ij}^* defined as

$$\frac{1}{m_{ij}^*} \equiv \frac{1}{\hbar^2}\frac{\partial^2 E(k)}{\partial k_i \partial k_j}. \tag{9}$$

The effective masses are listed in Appendix F for important semiconductors.

Carriers in motion are also characterized by a group velocity

$$v_g = \frac{1}{\hbar}\frac{dE}{dk} \tag{10}$$

and with momentum

$$p = \hbar k. \tag{11}$$

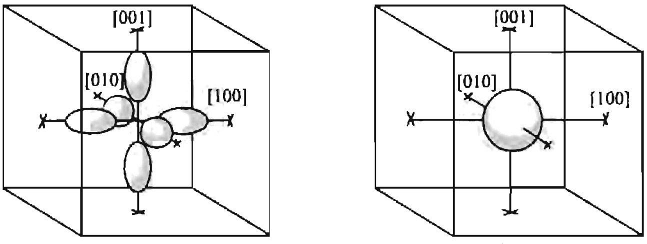
The conduction band consists of a number of subbands (Fig. 4). The bottom of the conduction band can appear at the center $k = 0$ (Γ) or off center along different k axes. Symmetry considerations alone do not determine the location of the bottom of the conduction band. Experimental results show, however, that in Si it is off center and along the [100] axis (Δ), and in GaAs the bottom is at $k = 0$ (Γ). Considering that the valence-band maximum (E_V) occurs at Γ, the conduction-band minimum can be aligned or misaligned in k-space in determining the bandgap. This results in direct bandgap for GaAs and indirect bandgap for Si. This bears significant consequences when carriers transfer between this minimum gap in that momentum (or k) is conserved for direct bandgap but changed for indirect bandgap.

Figure 5 shows the shapes of the constant-energy surfaces. For Si there are six ellipsoids along the $\langle 100 \rangle$-axes, with the centers of the ellipsoids located at about three-fourths of the distance from the Brillouin zone center. For GaAs the constant energy surface is a sphere at the zone center. By fitting experimental results to parabolic bands, we obtain the electron effective masses; one for GaAs and two for Si, m_l^* along the symmetry axes and m_t^* transverse to the symmetry axes. Appendix G also includes these values.

At room temperature and under normal atmospheric pressure, the values of the bandgap are 1.12 eV for Si and 1.42 eV for GaAs. These values are for high-purity materials. For highly doped materials the bandgaps become smaller. Experimental results show that the bandgaps of most semiconductors decrease with increasing temperature. Figure 6 shows variations of bandgaps as a function of temperature for Si and GaAs. The bandgap approaches 1.17 and 1.52 eV respectively for these two semiconductors at 0 K. The variation of bandgaps with temperature can be expressed approximately by a universal function

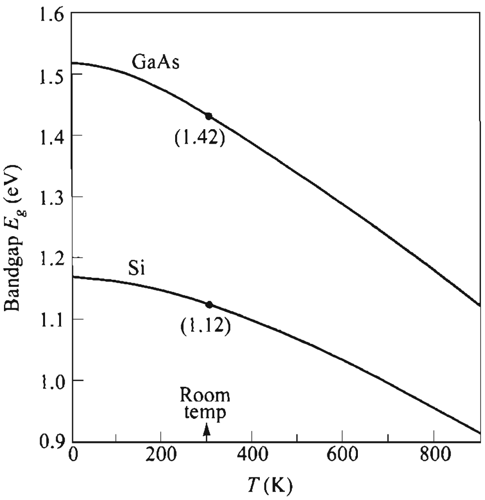
$$E_g(T) \approx E_g(0) - \frac{\alpha T^2}{T + \beta} \tag{12}$$

where $E_g(0)$, α, and β are given in the inset of Fig. 6. The temperature coefficient dE_g/dT is negative for both semiconductors. Some semiconductors have positive dE_g/dT; for example, the bandgap of PbS (Appendix F) increases from 0.286 eV at 0 K to 0.41 eV at 300 K. Near room temperature, the bandgap of GaAs increases with pressure P,[24] and dE_g/dP is about 12.6×10^{-6} eV-cm^2/N, while the Si bandgap decreases with pressure, with $dE_g/dP = -2.4 \times 10^{-6}$ eV-cm^2/N.



Si GaAs

Fig. 5 Shapes of constant-energy surfaces for electrons in Si and GaAs. For Si there are six ellipsoids along the $\langle 100 \rangle$-axes with the centers of the ellipsoids located at about three-fourths of the distance from the Brillouin zone center. For GaAs the constant-energy surface is a sphere at zone center. (After Ref. 21.)



	$E_g(0)$ (eV)	α (eV/K)	β (K)
GaAs	1.519	5.4×10^{-4}	204
Si	1.169	4.9×10^{-4}	655

$$E_g(T) = E_g(0) - \frac{\alpha T^2}{T+\beta}$$

Fig. 6 Energy bandgaps of Si and GaAs as a function of temperature. (After Refs. 22–23.)

1.4 CARRIER CONCENTRATION AT THERMAL EQUILIBRIUM

One of the most-important properties of a semiconductor is that it can be doped with different types and concentrations of impurities to vary its resistivity. Also, when these impurities are ionized and the carriers are depleted, they leave behind a charge density that results in an electric field and sometimes a potential barrier inside the semiconductor. Such properties are absent in a metal or an insulator.

Figure 7 shows three basic bond representations of a semiconductor. Figure 7a shows intrinsic silicon, which is very pure and contains a negligibly small amount of impurities. Each silicon atom shares its four valence electrons with the four neigh-

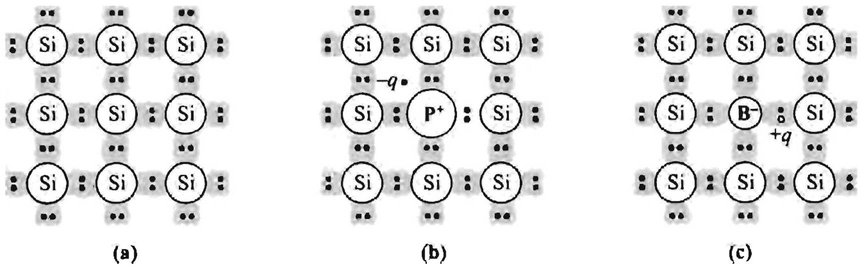


Fig. 7 Three basic bond pictures of a semiconductor. (a) Intrinsic Si with no impurity. (b) *n*-type Si with donor (phosphorus). (c) *p*-type Si with acceptor (boron).

boring atoms, forming four covalent bonds (also see Fig. 1). Figure 7b shows an *n*-type silicon, where a substitutional phosphorous atom with five valence electrons has replaced a silicon atom, and a *negative*-charged electron is *donated* to the lattice in the conduction band. The phosphorous atom is called a *donor*. Figure 7c similarly shows that when a boron atom with three valence electrons substitutes for a silicon atom, a *positive*-charged *hole* is created in the valence band, and an additional electron will be *accepted* to form four covalent bonds around the boron. This is *p*-type, and the boron is an *acceptor*.

These names of *n*- and *p*-type had been coined when it was observed that if a metal whisker was pressed against a *p*-type material, forming a Schottky barrier diode (see Chapter 3), a *positive* bias was required on the semiconductor to produce a noticeable current.[25,26] Also, when exposed to light, a *positive* potential was generated with respect to the metal whisker. Conversely, a *negative* bias was required on an *n*-type material to produce a large current.

1.4.1 Carrier Concentration and Fermi Level

We first consider the intrinsic case without impurities added to the semiconductor. The number of electrons (occupied conduction-band levels) is given by the total number of states $N(E)$ multiplied by the occupancy $F(E)$, integrated over the conduction band,

$$n = \int_{E_C}^{\infty} N(E)F(E)dE\,. \tag{13}$$

The density of states $N(E)$ can be approximated by the density near the bottom of the conduction band for low-enough carrier densities and temperatures:[5]

$$N(E) = M_C \frac{\sqrt{2}}{\pi^2}\frac{m_{de}^{3/2}(E-E_C)^{1/2}}{\hbar^3}\,. \tag{14}$$

M_C is the number of equivalent minima in the conduction band and m_{de} is the density-of-state effective mass for electrons:[5]

$$m_{de} = (m_1^* m_2^* m_3^*)^{1/3} \tag{15}$$

where m_1^*, m_2^*, m_3^* are the effective masses along the principal axes of the ellipsoidal energy surface. For example, in silicon $m_{de} = (m_l^* m_t^{*2})^{1/3}$. The occupancy is a strong function of temperature and energy, and is represented by the Fermi-Dirac distribution function

$$F(E) = \frac{1}{1+\exp[(E-E_F)/kT]} \tag{16}$$

where E_F is the Fermi energy level which can be determined from the charge neutrality condition (see Section 1.4.3).

The integral of Eq. 13 can be evaluated to be

$$n = N_C \frac{2}{\sqrt{\pi}} F_{1/2}\left(\frac{E_F - E_C}{kT}\right) \tag{17}$$

where N_C is the effective density of states in the conduction band and is given by

$$N_C \equiv 2\left(\frac{2\pi m_{de} kT}{h^2}\right)^{3/2} M_C . \tag{18}$$

The Fermi-Dirac integral, changing variables with $\eta \equiv (E - E_C)/kT$ and $\eta_F \equiv (E_F - E_C)/kT$, is given by

$$F_{1/2}\left(\frac{E_F - E_C}{kT}\right) \equiv F_{1/2}(\eta_F) = \int_{E_C}^{\infty} \frac{[(E - E_C)/kT]^{1/2}}{1 + \exp[(E - E_F)/kT]} \frac{dE}{kT}$$

$$= \int_0^{\infty} \frac{\eta^{1/2}}{1 + \exp(\eta - \eta_F)} d\eta \tag{19}$$

whose values are plotted in Fig. 8. Note that for $\eta_F < -1$, the integral can be approximated by an exponential function. At $\eta_F = 0$ when the Fermi level coincides with the band edge, the integral has a value of ≈ 0.6 such that $n \approx 0.7 N_C$.

Nondegenerate Semiconductors. By definition, in nondegenerate semiconductors, the doping concentrations are smaller than N_C and the Fermi levels are more than several kT below E_C (negative η_F), the Fermi-Dirac integral approaches

$$F_{1/2}\left(\frac{E_F - E_C}{kT}\right) = \frac{\sqrt{\pi}}{2} \exp\left(-\frac{E_C - E_F}{kT}\right) \tag{20}$$

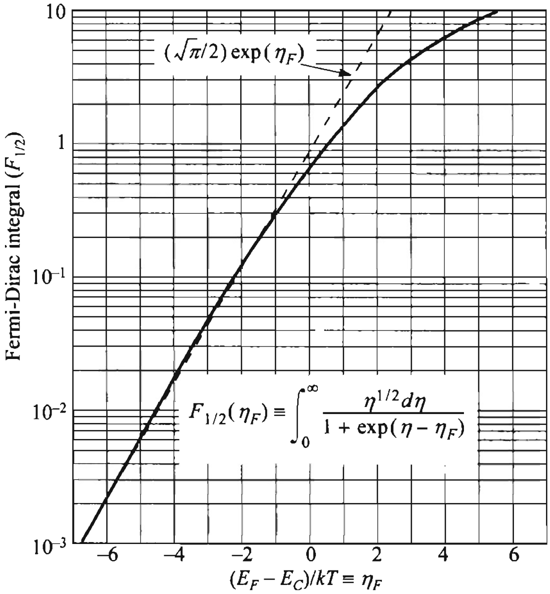


Fig. 8 Fermi-Dirac integral $F_{1/2}$ as a function of Fermi energy. (After Ref. 27.) Dashed line is approximation of Boltzmann statistics.

and Boltzmann statistics apply. Equation 17 becomes

$$n = N_C \exp\left(-\frac{E_C - E_F}{kT}\right) \quad \text{or} \quad E_C - E_F = kT\ln\left(\frac{N_C}{n}\right). \tag{21}$$

Similarly, for p-type semiconductors we can obtain the hole density and its Fermi level near the top of the valence band;

$$p = N_V \frac{2}{\sqrt{\pi}} F_{1/2}\left(\frac{E_V - E_F}{kT}\right) \tag{22}$$

which can be simplified to

$$p = N_V \exp\left(-\frac{E_F - E_V}{kT}\right) \quad \text{or} \quad E_F - E_V = kT\ln\left(\frac{N_V}{p}\right), \tag{23}$$

where N_V is the effective density of states in the valence band and is given by

$$N_V \equiv 2\left(\frac{2\pi m_{dh} kT}{h^2}\right)^{3/2}. \tag{24}$$

Here m_{dh} is the density-of-state effective mass of the valence band:[5]

$$m_{dh} = (m_{lh}^{*\,3/2} + m_{hh}^{*\,3/2})^{2/3} \tag{25}$$

where the subscripts refer to *light* and *heavy* hole masses previously referenced in Eq. 9.

Degenerate Semiconductors. As shown in Fig. 8, for degenerate levels where n- or p-concentrations are near or beyond the effective density of states (N_C or N_V), the value of Fermi-Dirac integral has to be used instead of the simplified Boltzmann statistics. For $\eta_F > -1$, the integral has weaker dependence on the carrier concentration. Note that also the Fermi levels are outside the energy gap. A useful estimate of the Fermi level as a function of carrier concentration is given by, for n-type semiconductor[28]

$$E_F - E_C \approx kT\left[\ln\left(\frac{n}{N_C}\right) + 2^{-3/2}\left(\frac{n}{N_C}\right)\right], \tag{26a}$$

and for p-type

$$E_V - E_F \approx kT\left[\ln\left(\frac{p}{N_V}\right) + 2^{-3/2}\left(\frac{p}{N_V}\right)\right]. \tag{26b}$$

Intrinsic Concentration. For intrinsic semiconductors at finite temperatures, thermal agitation occurs which results in continuous excitation of electrons from the valence band to the conduction band, and leaving an equal number of holes in the valence band. This process is balanced by recombination of the electrons in the conduction band with holes in the valence band. At steady state, the net result is $n = p = n_i$, where n_i is the intrinsic carrier density.

The Fermi level for an intrinsic semiconductor (which by definition is nondegenerate) is obtained by equating Eqs. 21 and 23:

$$E_F = E_i = \frac{E_C + E_V}{2} + \frac{kT}{2}\ln\left(\frac{N_V}{N_C}\right)$$
$$= \frac{E_C + E_V}{2} + \frac{3kT}{4}\ln\left(\frac{m_{dh}}{m_{de}M_C^{2/3}}\right) . \qquad (27)$$

Hence the Fermi level E_i of an intrinsic semiconductor generally lies very close to, but not exactly at, the middle of the bandgap. The intrinsic carrier density n_i can be obtained from Eq. 21 or 23:

$$n_i = N_C\exp\left(-\frac{E_C - E_i}{kT}\right) = N_V\exp\left(-\frac{E_i - E_V}{kT}\right) = \sqrt{N_C N_V}\exp\left(-\frac{E_g}{2kT}\right)$$
$$= 4.9\times10^{15}\left(\frac{m_{de}m_{dh}}{m_0^2}\right)^{3/4} M_C^{1/2}T^{3/2}\exp\left(-\frac{E_g}{2kT}\right) . \qquad (28)$$

Figure 9 shows the temperature dependence of n_i for Si and GaAs. As expected, the larger the bandgap is, the smaller the intrinsic carrier density will be.[30]

It also follows that for nondegenerate semiconductors, the product of the majority and minority carrier concentrations is fixed to be

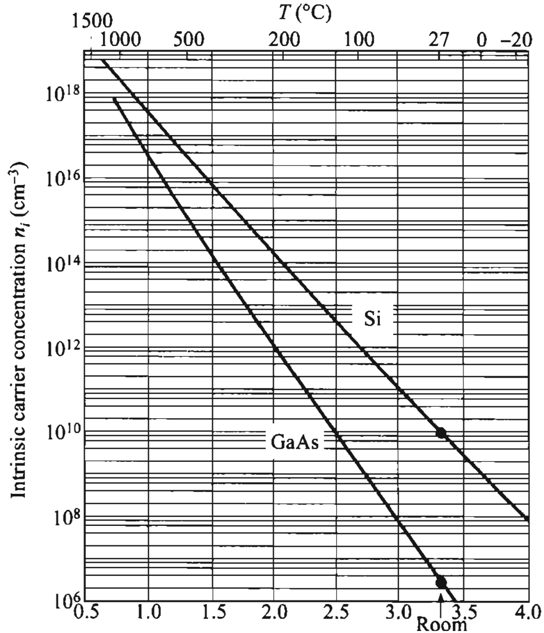


Fig. 9 Intrinsic carrier concentrations of Si and GaAs as a function of reciprocal temperature. (After Refs. 22 and 29.)

$$pn = N_C N_V \exp\left(-\frac{E_g}{kT}\right)$$
$$= n_i^2 \ , \qquad (29)$$

which is known as the mass-action law. But for degenerate semiconductors, $pn < n_i^2$. Also using Eq. 28 and E_i as the reference energy, we have the alternate equations for n-type materials;

$$n = n_i \exp\left(\frac{E_F - E_i}{kT}\right) \quad \text{or} \quad E_F - E_i = kT\ln\left(\frac{n}{n_i}\right) \ , \qquad (30a)$$

and for p-type materials;

$$p = n_i \exp\left(\frac{E_i - E_F}{kT}\right) \quad \text{or} \quad E_i - E_F = kT\ln\left(\frac{p}{n_i}\right) . \qquad (30b)$$

1.4.2 Donors and Acceptors

When a semiconductor is doped with donor or acceptor impurities, impurity energy levels are introduced that usually lie within the energy gap. A donor impurity has a donor level which is defined as being neutral if filled by an electron, and positive if empty. Conversely, an acceptor level is neutral if empty and negative if filled by an electron. These energy levels are important in calculating the fraction of dopants being ionized, or electrically active, as discussed in Section 1.4.3.

To get a feeling of the magnitude of the impurity ionization energy, we use the simplest calculation based on the hydrogen-atom model. The ionization energy for the hydrogen atom in vacuum is

$$E_H = \frac{m_0 q^4}{32\pi^2 \varepsilon_0^2 \hbar^2} = 13.6 \text{ eV} . \qquad (31)$$

The ionization energy for a donor $(E_C - E_D)$ in a lattice can be obtained by replacing m_0 by the conductivity effective mass of electrons[5]

$$m_{ce} = 3\left(\frac{1}{m_1^*} + \frac{1}{m_2^*} + \frac{1}{m_3^*}\right)^{-1} \qquad (32)$$

and by replacing ε_0 by the permittivity of the semiconductor ε_s in Eq. 31:

$$E_C - E_D = \left(\frac{\varepsilon_0}{\varepsilon_s}\right)^2 \left(\frac{m_{ce}}{m_0}\right) E_H . \qquad (33)$$

The ionization energy for donors as calculated from Eq. 33 is 0.025 eV for Si and 0.007 eV for GaAs. The hydrogen-atom calculation for the ionization level for the acceptors is similar to that for the donors. The calculated acceptor ionization energy (measured from the valence-band edge, $E_a \equiv (E_A - E_V)$ is 0.05 eV for Si and GaAs.

Although this simple hydrogen-atom model given above certainly cannot account for the details of ionization energy, particularly the deep levels in semiconductors,[31-33] the calculated values do predict the correct order of magnitude of the true ionization energies for shallow impurities. These calculated values are shown to be much smaller than the energy gap, and often are referred to as shallow impurities if they are close to the band edges. Also, since these small ionization energies are comparable to the thermal energy kT, ionization is usually complete at room temperature. Figure 10 shows the measured ionization energies for various impurities in Si and GaAs. Note that it is possible for a single atom to have many levels; for example, gold in silicon has both an acceptor level and a donor level in the forbidden energy gap.

1.4.3 Calculation of Fermi Level

The Fermi level for the intrinsic semiconductor (Eq. 27) lies very close to the middle of the bandgap. Figure 11a depicts this situation, showing schematically from left to right the simplified band diagram, the density of states $N(E)$, the Fermi-Dirac distribution function $F(E)$, and the carrier concentrations. The shaded areas in the conduction band and the valence band represent electrons and holes, and their numbers are the same; i.e., $n = p = n_i$ for the intrinsic case.

When impurities are introduced to the semiconductor crystals, depending on the impurity energy level and the lattice temperature, not all dopants are necessarily ionized. The ionized concentration for donors is given by[36]

$$N_D^+ = \frac{N_D}{1 + g_D \exp[(E_F - E_D)/kT]} \tag{34}$$

where g_D is the ground-state degeneracy of the donor impurity level and equal to 2 because a donor level can accept one electron with either spin (or can have no electron). When acceptor impurities of concentration N_A are added to a semiconductor crystal, a similar expression can be written for the ionized acceptors

$$N_A^- = \frac{N_A}{1 + g_A \exp[(E_A - E_F)/kT]} \tag{35}$$

where the ground-state degeneracy factor g_A is 4 for acceptor levels. The value is 4 because in most semiconductors each acceptor impurity level can accept one hole of either spin and the impurity level is doubly degenerate as a result of the two degenerate valence bands at $\boldsymbol{k} = 0$.

When impurity atoms are introduced, the total negative charges (electrons and ionized acceptors) must equal the total positive charges (holes and ionized donors), represented by the charge neutrality

$$n + N_A^- = p + N_D^+ . \tag{36}$$

With impurities added, the mass-action law ($pn = n_i^2$) in Eq. 29 still applies (until degeneracy), and the pn product is always independent of the added impurities.

Consider the case shown in Fig. 11b, where donor impurities with the concentration N_D (cm^{-3}) are added to the crystal. The charge neutrality condition becomes

$$\begin{aligned} n &= N_D^+ + p \\ &\approx N_D^+ \quad . \end{aligned} \tag{37}$$

With substitution, we obtain

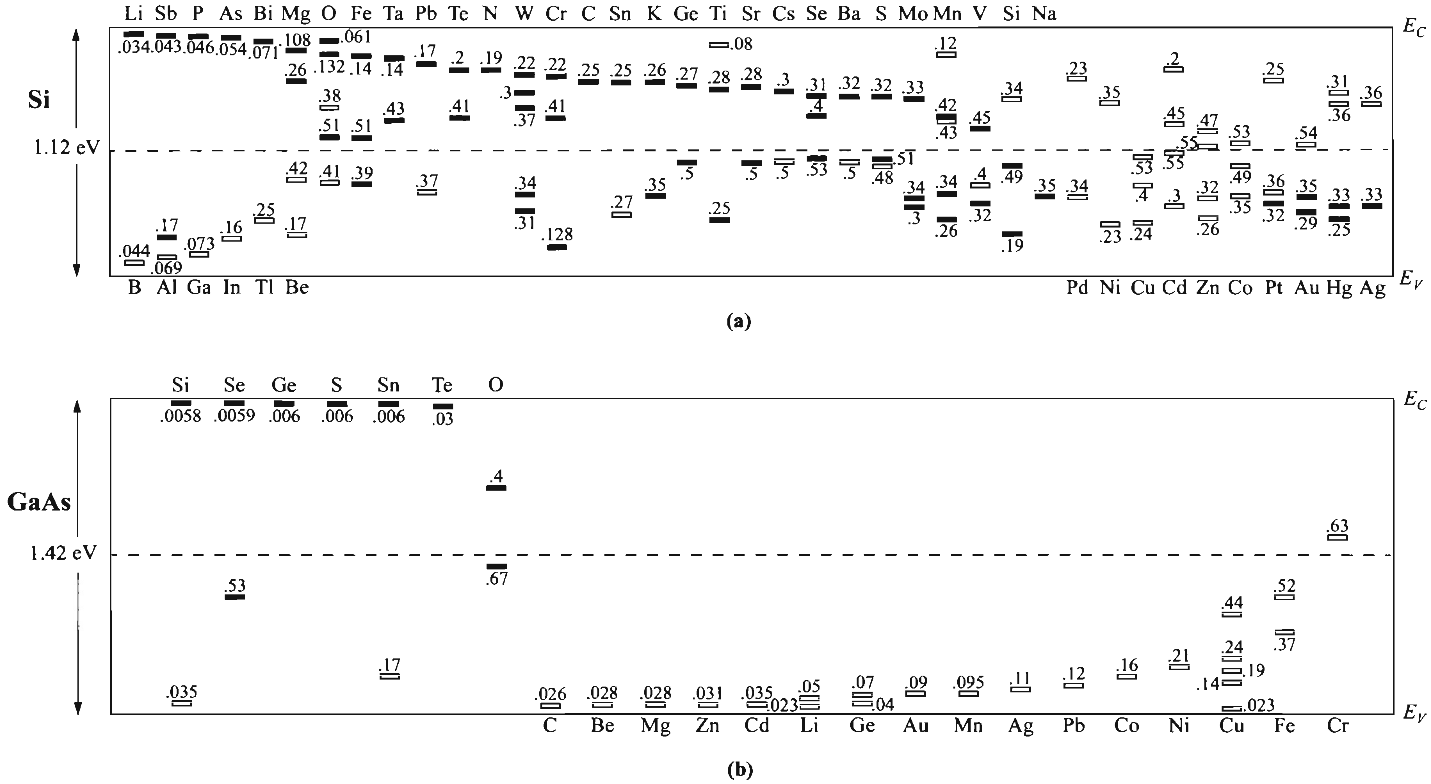


Fig. 10 Measured ionization energies for varies impurities in (a) Si and (b) GaAs. Levels below the gap center are measured from E_V. Levels above the gap center are measured from E_C. Solid bars represent donor levels and hollow boxes represent acceptor levels. (After Refs. 29, 31, 34, and 35.)

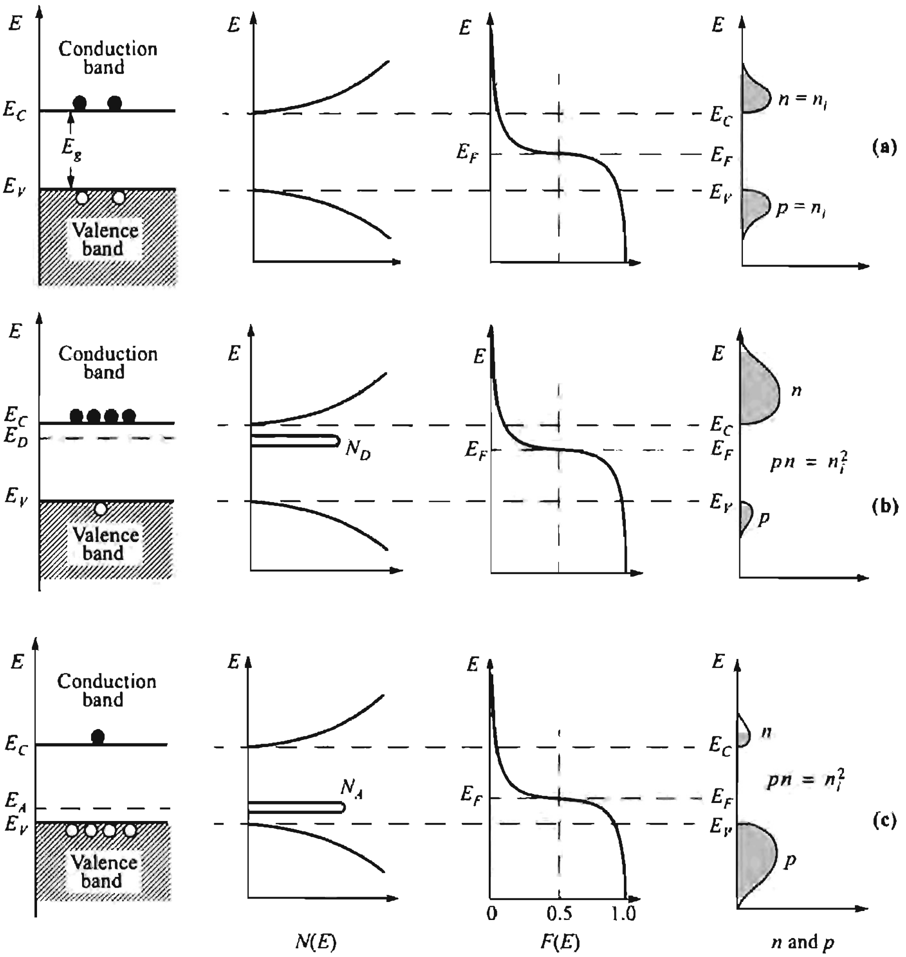

E
Conduction band
E_C
E_g
E_V
Valence band
E_F
$n = n_i$
$p = n_i$
(a)
E_D
N_D
n
$pn = n_i^2$
p
(b)
E_A
N_A
(c)
0
0.5
1.0
N(E)
F(E)
n and p

Fig. 11 Schematic band diagram, density of states, Fermi-Dirac distribution, and carrier concentrations for (a) intrinsic, (b) n-type, and (c) p-type semiconductors at thermal equilibrium. Note that $pn = n_i^2$ for all three cases.

$$N_C \exp\left(-\frac{E_C - E_F}{kT}\right) \approx \frac{N_D}{1 + 2\exp[(E_F - E_D)/kT]}. \tag{38}$$

Thus for a set of given N_D, E_D, N_C, and T, the Fermi level E_F can be uniquely determined implicitly. Knowing E_F, the carrier concentrations n can be calculated. Equation 38 can also be solved graphically. In Fig. 12, the values of n and N_D^+ are plotted as a function of E_F. Where the two curves meet determines the position of E_F.

Without solving for Eq. 38, it can be shown that for $N_D \gg \frac{1}{2}N_C\exp[-(E_C-E_D)/kT] \gg N_A$, the electron concentration can be approximated by[5]

$$n \approx \sqrt{\frac{N_D N_C}{2}}\exp\left[-\frac{(E_C-E_D)}{2kT}\right]. \tag{39}$$

For compensated n-type material ($N_D > N_A$) with nonnegligible acceptor concentration, when $N_A \gg \frac{1}{2}N_C\exp[-(E_C-E_D)/kT]$, the approximate expression for the electron density is then

$$n \approx \left(\frac{N_D-N_A}{2N_A}\right)N_C\exp\left[-\frac{(E_C-E_D)}{kT}\right]. \tag{40}$$

Figure 13 shows a typical example, where n is plotted as a function of the reciprocal temperature. At high temperatures we have the intrinsic range since $n \approx p \approx n_i \gg N_D$. At medium temperatures, $n \approx N_D$. At very low temperatures most impurities are frozen out and the slope is given by either Eq. 39 or Eq. 40, depending on the compensation conditions. The electron density, however, remains essentially constant over a wide range of temperatures (≈100 to 500 K).

Figure 14 shows the Fermi level for Si and GaAs as a function of temperature and impurity concentration, as well as the dependence of the bandgap on temperature (see Fig. 6).

At relatively high temperatures, most donors and acceptors are ionized, so the neutrality condition can be approximated by

$$n + N_A = p + N_D. \tag{41}$$

Equations 29 and 41 can be combined to give the concentrations of electrons and holes. In an n-type semiconductor where $N_D > N_A$:

$$\begin{aligned} n_{no} &= \frac{1}{2}\left[(N_D-N_A)+\sqrt{(N_D-N_A)^2+4n_i^2}\right] \\ &\approx N_D \qquad \text{if } |N_D-N_A| \gg n_i \text{ or } N_D \gg N_A, \end{aligned} \tag{42}$$

$$p_{no} = \frac{n_i^2}{n_{no}} \approx \frac{n_i^2}{N_D}. \tag{43}$$

The Fermi level can be obtained from

$$n_{no} = N_D = N_C\exp\left(-\frac{E_C-E_F}{kT}\right) = n_i\exp\left(\frac{E_F-E_i}{kT}\right). \tag{44}$$

Similarly, the carrier concentrations in a p-type semiconductor ($N_A > N_D$) are given by

$$\begin{aligned} p_{po} &= \frac{1}{2}\left[(N_A-N_D)+\sqrt{(N_A-N_D)^2+4n_i^2}\right] \\ &\approx N_A \qquad \text{if } |N_A-N_D| \gg n_i \text{ or } N_A \gg N_D, \end{aligned} \tag{45}$$

$$n_{po} = \frac{n_i^2}{p_{po}} \approx \frac{n_i^2}{N_A}, \tag{46}$$

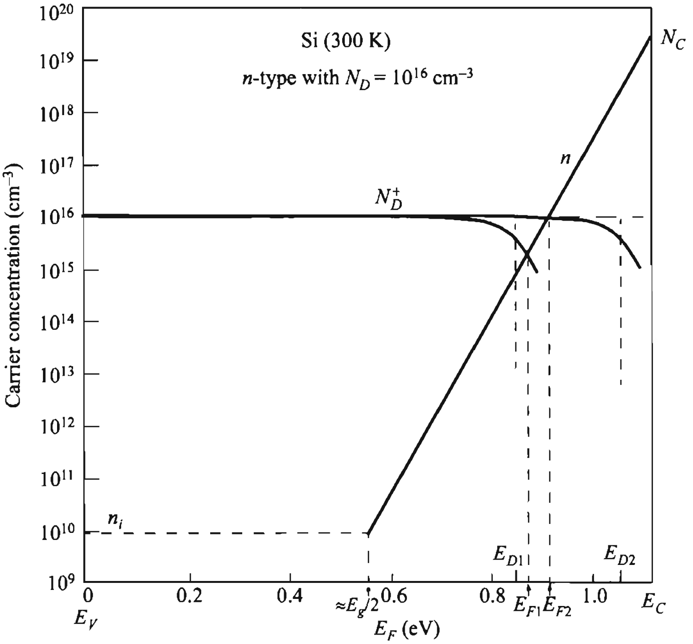


Fig. 12 Graphical method to determine the Fermi energy level E_F and electron concentration n, when ionization is not complete. Examples with two different values of impurity levels E_D are shown.

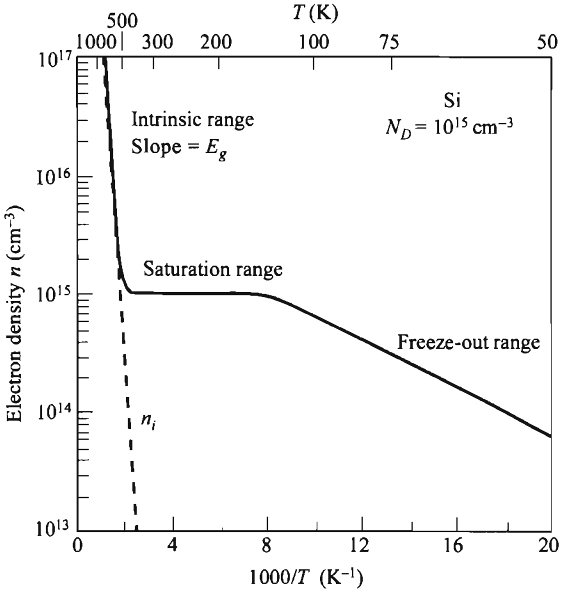


Fig. 13 Electron density as a function of temperature for a Si sample with donor impurity concentration of 10^{15} cm^{-3}. (After Ref. 5.)

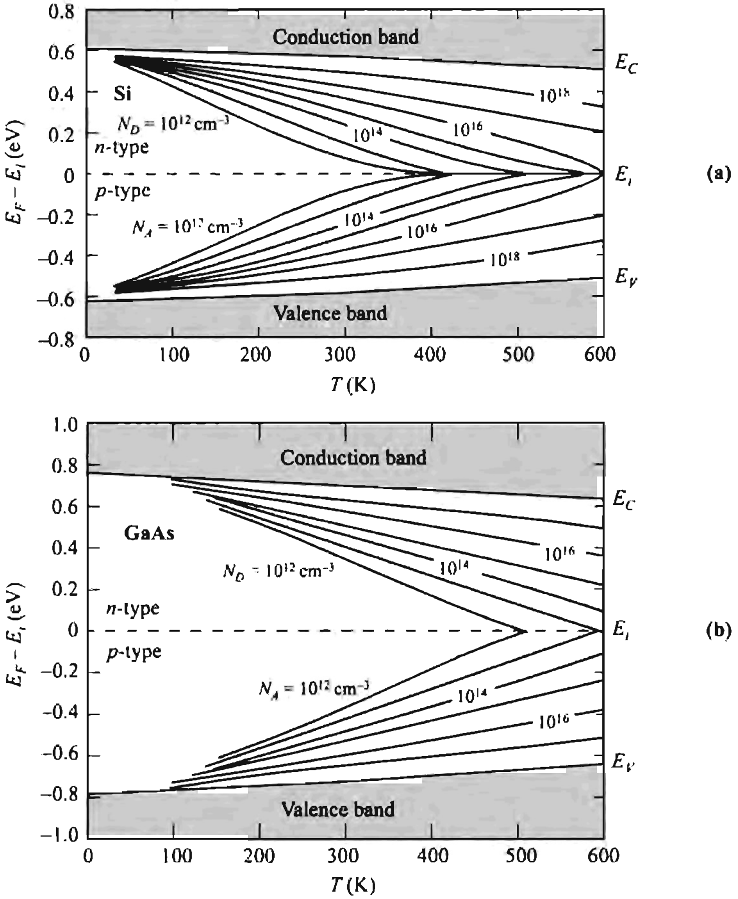


Fig. 14 Fermi level for (a) Si and (b) GaAs as a function of temperature and impurity concentration. The dependence of the bandgap on temperature is also shown. (After Ref. 37.)

and

$$p_{po} = N_A = N_V \exp\left(-\frac{E_F - E_V}{kT}\right) = n_i \exp\left(\frac{E_i - E_F}{kT}\right). \tag{47}$$

In the formulas above, the subscripts n and p refer to the type of semiconductors, and the subscript "o" refers to the thermal equilibrium condition. For n-type semiconductors the electron is referred to as the majority carrier and the hole as the minority carrier, since the electron concentration is the larger of the two. The roles are reversed for p-type semiconductors.

1.5 CARRIER-TRANSPORT PHENOMENA

1.5.1 Drift and Mobility

At low electric fields, the drift velocity v_d is proportional to the electric field strength $\mathscr{E}$ and the proportionality constant is defined as the mobility μ in cm^2/V-s, or

$$v_d = \mu\mathscr{E}. \tag{48}$$

For nonpolar semiconductors, such as Ge and Si, the presence of acoustic phonons (see Section 1.6.1) and ionized impurities results in carrier scattering that significantly affects the mobility. The mobility from interaction with acoustic phonon of the lattice, μ_l, is given by[38]

$$\mu_l = \frac{\sqrt{8\pi}\, q\hbar^4 C_l}{3E_{ds}^2 m_c^{*5/2}(kT)^{3/2}} \propto \frac{1}{m_c^{*5/2}T^{3/2}} \tag{49}$$

where C_l is the average longitudinal elastic constant of the semiconductor, E_{ds} the displacement of the band edge per unit dilation of the lattice, and m_c^* the conductivity effective mass. From Eq. 49 mobility decreases with the temperature and with the effective mass.

The mobility from ionized impurities μ_i can be described by[39]

$$\mu_i = \frac{64\sqrt{\pi}\,\varepsilon_s^2(2kT)^{3/2}}{N_I q^3 m^{*1/2}}\left\{\ln\left[1+\left(\frac{12\pi\varepsilon_s kT}{q^2 N_I^{1/3}}\right)^2\right]\right\}^{-1} \propto \frac{T^{3/2}}{N_I m^{*1/2}} \tag{50}$$

where N_I is the ionized impurity density. The mobility is expected to decrease with the effective mass but to increase with the temperature because carriers with higher thermal velocity are less deflected by Coulomb scattering. Note the common dependence of the two scattering events on the effective mass but opposite dependence on temperature. The combined mobility, which includes the two mechanisms above, is given by the Matthiessen rule

$$\mu = \left(\frac{1}{\mu_l}+\frac{1}{\mu_i}\right)^{-1}. \tag{51}$$

In addition to the scattering mechanisms discussed above, other mechanisms also affect the actual mobility. For example, (1) the intravalley scattering in which an electron is scattered within an energy ellipsoid (Fig. 5) and only long-wavelength phonons (acoustic phonons) are involved; and (2) the intervalley scattering in which an electron is scattered from the vicinity of one minimum to another minimum and an energetic phonon (optical phonon) is involved. For polar semiconductors such as GaAs, polar-optical-phonon scattering is significant.

Qualitatively, since mobility is controlled by scattering, it can also be related to the mean free time τ_m or mean free path λ_m by

$$\mu = \frac{q\tau_m}{m^*} = \frac{q\lambda_m}{\sqrt{3kTm^*}}. \tag{52}$$

The last term uses the relationship

$$\lambda_m = v_{th}\tau_m \tag{53}$$

where v_{th} is the thermal velocity given by

$$v_{th} = \sqrt{\frac{3kT}{m^*}}. \tag{54}$$

For multiple scattering mechanisms, the effective mean free time is derived from the individual mean free times of scattering events by

$$\frac{1}{\tau_m} = \frac{1}{\tau_{m1}} + \frac{1}{\tau_{m2}} + \dots \tag{55}$$

It can be seen that Eqs. 51 and 55 are equivalent.

Figure 15 shows the measured mobilities of Si and GaAs versus impurity concentrations at room temperature. As the impurity concentration increases (at room temperature most shallow impurities are ionized) the mobility decreases, as predicted by Eq. 50. Also for larger m^*, μ decreases; thus for a given impurity concentration the electron mobilities for these semiconductors are larger than the hole mobilities (Appendixes F and G list the effective masses).

Figure 16 shows the temperature effect on mobility for n-type and p-type silicon samples. For lower impurity concentrations the mobility is limited by phonon scattering and it decreases with temperature as predicted by Eq. 49. The measured slopes, however, are different from –3/2 because of other scattering mechanisms. For these

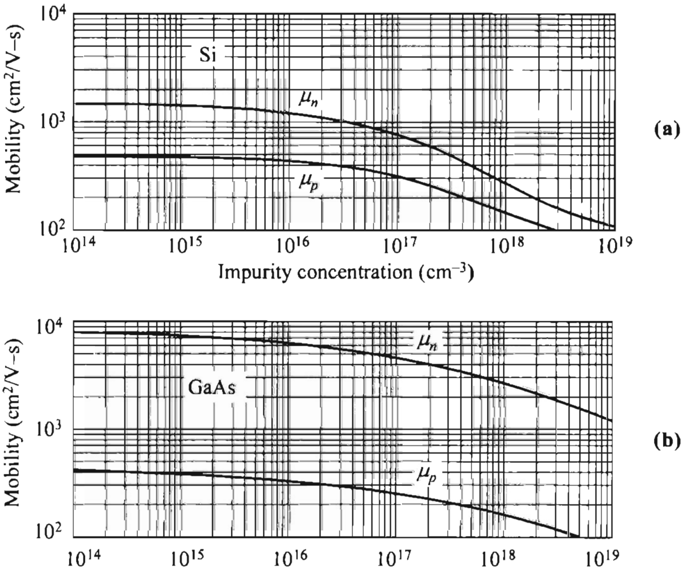


Fig. 15 Drift mobility of (a) Si (After Ref. 40.) and (b) GaAs at 300 K vs. impurity concentration (after Ref. 11).

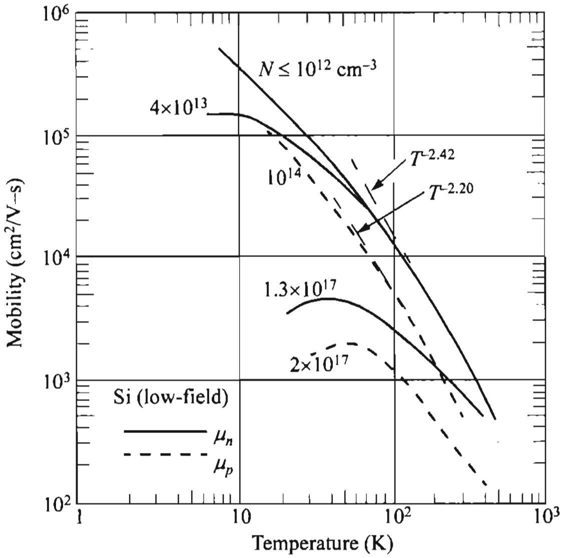


Fig. 16 Mobility of electrons and holes in Si as a function of temperature. (After Ref. 41.)

pure materials, near room temperature, the mobility varies as $T^{-2.42}$ and $T^{-2.20}$ for n- and p-type Si, respectively; and as $T^{-1.0}$ and $T^{-2.1}$ for n- and p-type GaAs (not shown), respectively.

The mobilities discussed above are the conductivity mobilities, which have been shown to be equal to the drift mobilities.[34] They are, however, different from but related to the Hall mobilities considered in the next section.

1.5.2 Resistivity and Hall Effect

For semiconductors with both electrons and holes as carriers, the drift current under an applied field is given by

$$\begin{aligned} J &= \sigma\mathscr{E} \\ &= q(\mu_n n + \mu_p p)\mathscr{E} \end{aligned} \tag{56}$$

where σ is the conductance

$$\sigma = \frac{1}{\rho} = q(\mu_n n + \mu_p p) \tag{57}$$

and ρ is the resistivity. If $n \gg p$, as in n-type semiconductors,

$$\rho = \frac{1}{q\mu_n n} \tag{58}$$

and

$$\sigma = q\mu_n n . \tag{59}$$

The most-common method for measuring resistivity is the four-point probe method (insert, Fig. 17),[42,43] A small constant current is passed through the outer two probes and the voltage is measured between the inner two probes. For a thin wafer with thickness W much smaller than either a or d, the sheet resistance $R_\square$ is given by

$$R_\square = \frac{V}{I} \cdot \mathrm{CF} \qquad \Omega/\square \tag{60}$$

where CF is the correction factor shown in Fig. 17. The resistivity is then

$$\rho = R_\square W \qquad \Omega\text{-cm}. \tag{61}$$

In the limit when $d \gg S$, where S is the probe spacing, the correction factor becomes $\pi/\ln 2$ (= 4.54).

Figure 18a shows the measured resistivity (at 300 K) as a function of the impurity concentration (n-type phosphorus and p-type boron) for silicon. Resistivity is not a linear function of concentration because mobility is not constant and usually decreases with increasing concentration. Figure 18b shows the measured resistivities for GaAs. We can obtain the impurity concentration of a semiconductor if its resistivity is known and vice versa. Note that the impurity concentration may be different from the carrier concentration because of incomplete ionization. For example, in a p-type silicon with 10^{17} cm^{-3} gallium acceptor impurities, unionized acceptors at room temperature make up about 23% (from Eq. 35, Figs. 10 and 14); in other words, the carrier concentration is only 7.7×10^{16} cm^{-3}.

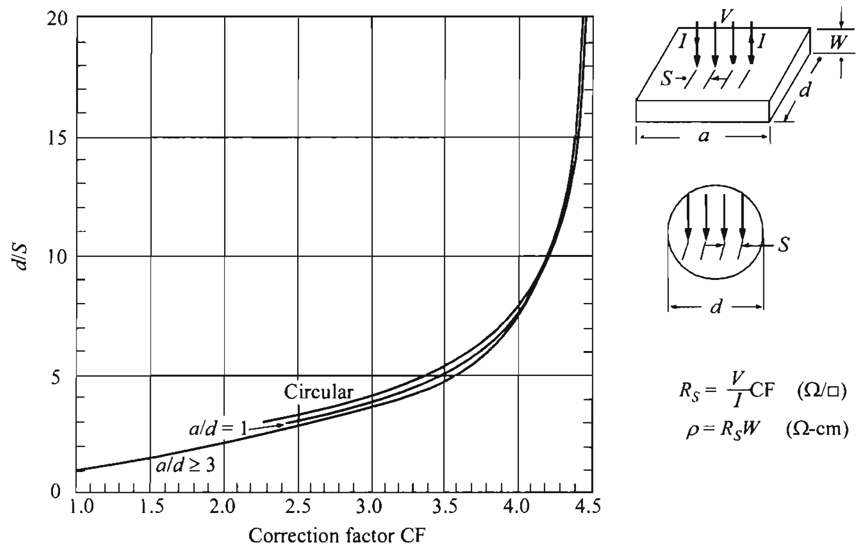


Fig. 17 Correction factor for measurement of resistivity using a four-point probe. (After Ref. 42.)

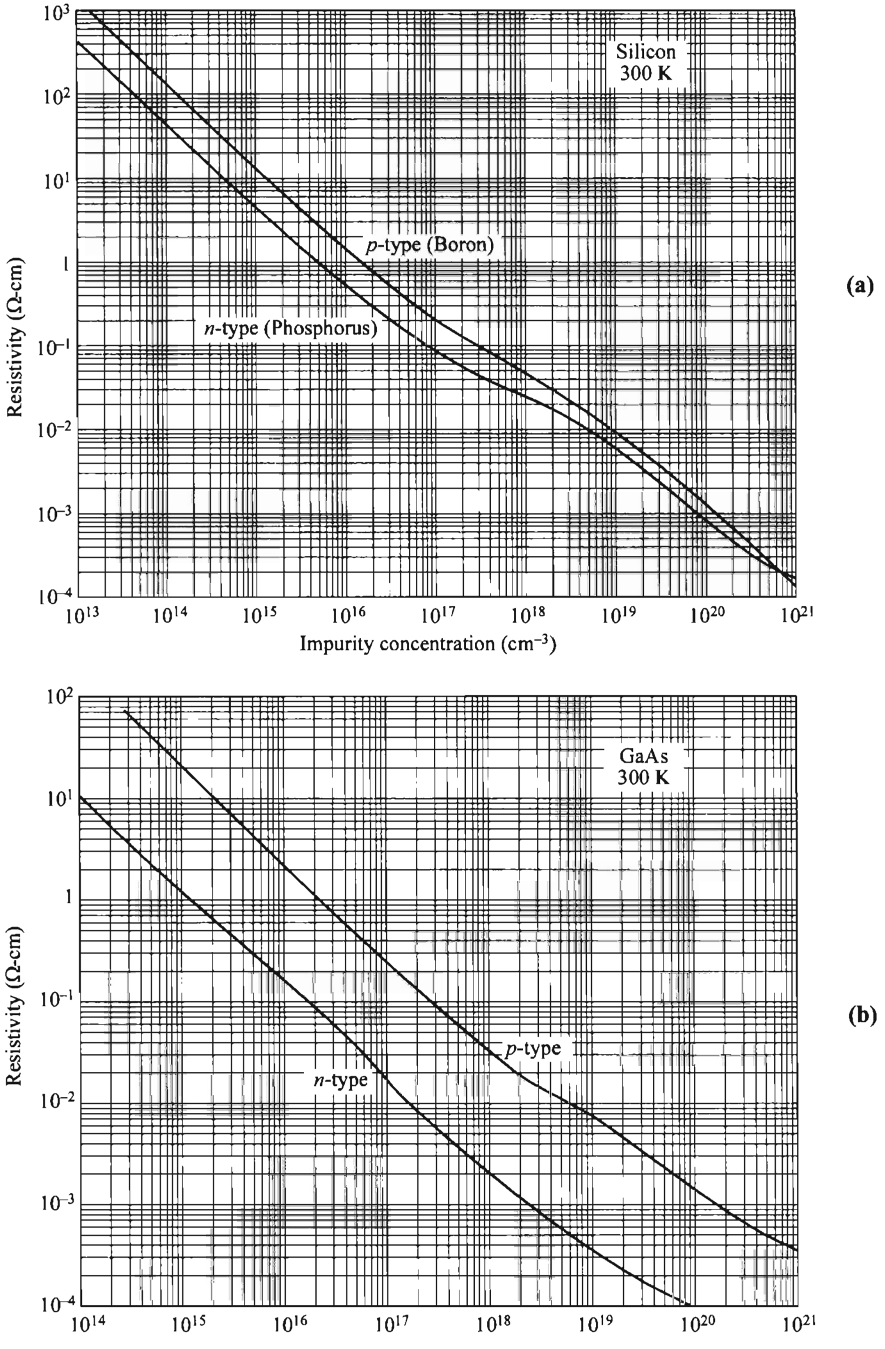


Fig. 18 Resistivity vs. impurity concentration at 300 K for (a) silicon (after Ref. 40) and (b) GaAs (after Ref. 35).

Hall Effect. Measurement of the resistivity only gives the product of the mobility and carrier concentration. To measure each parameter directly, the most-common method uses the Hall effect. The Hall effect is named after the scientist who made the discovery in 1879.[44] Even today it remains one of the most fascinating phenomena and is both fundamentally interesting and practical. Examples include the recent study of the fractional quantum Hall effect and the applications as magnetic-field sensors. The Hall effect is used in common practice to measure certain properties of semiconductors: namely, the carrier concentration (even down to a low level of 10^{12} cm^{-3}), the mobility, and the type (n or p). It is an important analytical tool since a simple conductance measurement can only give the product of concentration and mobility, and the type remains unknown.

Figure 19 shows the basic setup where an electric field is applied along the x-axis and a magnetic field is applied along the z-axis.[45] Consider a p-type sample. The Lorentz force exerts an average downward force on the holes

$$\text{Lorentz force} = qv_x \times \mathscr{B}_z, \tag{62}$$

and the downward-directed current causes a piling up of holes at the bottom side of the sample, which in turn gives rise to an electric field $\mathscr{E}_y$. Since there is no net current along the y-direction in the steady state, the electric field along the y-axis (Hall field) balances exactly the Lorentz force such that the carriers travel in a path parallel to the applied field $\mathscr{E}_x$. (For n-type material, electrons also pile up at the bottom surface, setting up a voltage of opposite polarity.)

The carrier velocity v is related to the current density by

$$J_x = qv_x p. \tag{63}$$

Since for each carrier the Lorentz force must be equal to the force exerted by the Hall field,

$$q\mathscr{E}_y = qv_x\mathscr{B}_z, \tag{64}$$

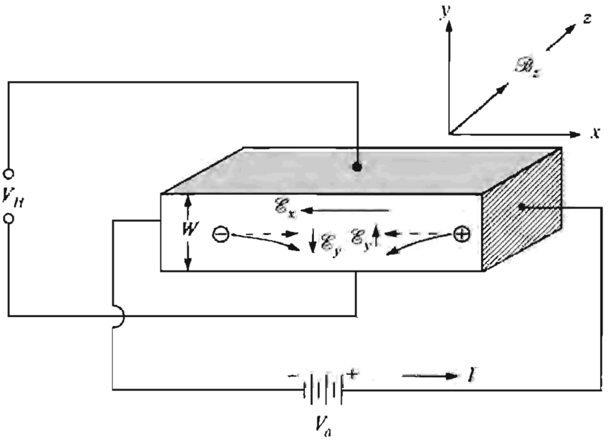


Fig. 19 Basic setup to measure carrier concentration using the Hall effect.

this Hall voltage can be measured externally and is given by

$$V_H = \mathscr{E}_y W = \frac{J_x \mathscr{B}_z W}{qp}. \tag{65}$$

When scattering is taken into account, the Hall voltage becomes

$$V_H = R_H J_x \mathscr{B}_z W \tag{66}$$

where R_H is the Hall coefficient and is given by

$$R_H = \frac{r_H}{qp} \qquad p \gg n, \tag{67a}$$

$$R_H = -\frac{r_H}{qn} \qquad n \gg p, \tag{67b}$$

with a Hall factor

$$r_H \equiv \frac{\langle \tau_m^2 \rangle}{\langle \tau_m \rangle^2}. \tag{68}$$

Thus, the carrier concentration and carrier type (electrons or holes from the polarity of the Hall voltage) can be obtained directly from the Hall measurement, provided that one type of carrier dominates and r_H is known.

Equation 67a or b also assumes conduction by a single type of carrier. A more-general solution is described by[5]

$$R_H = \frac{r_H}{q} \frac{\mu_p^2 p - \mu_n^2 n}{(\mu_p p + \mu_n n)^2}. \tag{69}$$

It can be seen in Eq. 69 that the sign of R_H and thus V_H reveals the majority type of the semiconductor sample.

The Hall mobility μ_H is defined as the product of the Hall coefficient and conductivity:

$$\mu_H = |R_H| \sigma. \tag{70}$$

The Hall mobility should be distinguished from the drift mobility μ_n (or μ_p) as given in Eq. 59 which does not contain the Hall factor r_H. Their relationship is given by

$$\mu_H = r_H \mu. \tag{71}$$

The parameter τ_m for the Hall factor is the mean free time between carrier collisions, which depends on the carrier energy. For example, for semiconductors with spherical constant-energy surfaces, $\tau_m \propto E^{-1/2}$ for phonon scattering and $\tau_m \propto E^{3/2}$ for ionized impurity scattering. In general,

$$\tau_m = C_1 E^{-s}, \tag{72}$$

where C_1 and s are constants. From Boltzmann distribution for nondegenerate semiconductors, the average value of the nth power of τ_m is

$$\langle \tau_m^n \rangle = \int_0^\infty \tau_m^n E^{3/2} \exp\left(-\frac{E}{kT}\right) dE \bigg/ \int_0^\infty E^{3/2} \exp\left(-\frac{E}{kT}\right) dE , \tag{73}$$

so that using the general form of τ_m, we obtain

$$\langle \tau_m^2 \rangle = \frac{C_1^2 (kT)^{-2s} \Gamma(\frac{5}{2} - 2s)}{\Gamma(\frac{5}{2})} \tag{74}$$

and

$$\langle \tau_m \rangle = \frac{C_1 (kT)^{-s} \Gamma(\frac{5}{2} - s)}{\Gamma(\frac{5}{2})} \tag{75}$$

where $\Gamma(n)$ is the gamma function defined as

$$\Gamma(n) \equiv \int_0^\infty x^{n-1} e^{-x} dx . \tag{76}$$

$[\Gamma(1/2) = \sqrt{\pi}.]$ From the expression above we obtain $r_H = 3\pi/8 = 1.18$ for phonon scattering and $r_H = 315\pi/512 = 1.93$ for ionized-impurity scattering. In general r_H lies in the range of 1–2. At very high magnetic fields, it approaches a value slightly below unity.

In the preceding discussion the applied magnetic field is assumed to be small enough that there is no change in the resistivity of the sample. However, under strong magnetic fields, a significant increase in the resistivity is observed, the so-called magnetoresistance effect, resulting from carriers travelling in a path that deviates from the applied electric field. For spherical-energy surfaces the ratio of the incremental resistivity to the bulk resistivity at zero magnetic field is given by[5]

$$\frac{\Delta\rho}{\rho_0} = \left\{ \left[\frac{\Gamma^2(\frac{5}{2})\Gamma(\frac{5}{2}-3s)}{\Gamma^3(\frac{5}{2}-s)} \right] \left(\frac{\mu_n^3 n + \mu_p^3 p}{\mu_n n + \mu_p p} \right) - \left[\frac{\Gamma(\frac{5}{2})\Gamma(\frac{5}{2}-2s)}{\Gamma^2(\frac{5}{2}-s)} \right]^2 \left(\frac{\mu_n^2 n - \mu_p^2 p}{\mu_n n + \mu_p p} \right)^2 \right\} \mathscr{B}_z^2 . \tag{77}$$

The ratio is proportional to the square of the magnetic field component perpendicular to the direction of the current flow. For $n \gg p$, $(\Delta\rho/\rho_0) \propto \mu_n^2 \mathscr{B}_z^2$. A similar result can be obtained for the case $p \gg n$.

1.5.3 High-Field Properties

In the preceding sections we considered the effect of low electric field on the transport of carriers in semiconductors. In this section we briefly consider some special effects and properties of semiconductors when the electric field is increased to moderate and high levels.

As discussed in Section 1.5.1, at low electric fields the drift velocity in a semiconductor is proportional to the field and the proportionality constant is the mobility that is independent of the electric field. When the fields are sufficiently large, however, nonlinearities in mobility and, in some cases, saturation of drift velocity are observed. At still larger fields, impact ionization occurs. First, we consider the nonlinear mobility.

At thermal equilibrium the carriers both emit and absorb phonons and the net rate of exchange of energy is zero. The energy distribution at thermal equilibrium is Maxwellian. In the presence of an electric field the carriers acquire energy from the field and lose it to phonons by emitting more phonons than are absorbed. At moderately high fields, the most frequent scattering events involve in the emission of acoustic phonons. Thus, the carriers on average acquire more energy than they have at thermal equilibrium. As the field increases, the average energy of the carriers also increases and they acquire an effective temperature T_e that is higher than the lattice temperature T. Balancing the rate at which energy is transferred from the field to the carriers by an equal rate of energy loss to the lattice, we obtain from the rate equation, for Ge and Si (semiconductors without transferred-electron effect):[3]

$$\frac{T_e}{T} = \frac{1}{2}\left[1 + \sqrt{1 + \frac{3\pi}{8}\left(\frac{\mu_0 \mathscr{E}}{c_s}\right)^2}\right] \tag{78}$$

and

$$v_d = \mu_0 \mathscr{E} \sqrt{\frac{T}{T_e}} \tag{79}$$

where μ_0 is the low-field mobility, and c_s the velocity of sound. For moderate field strength when $\mu_0 \mathscr{E}$ is comparable to c_s, the carrier velocity v_d starts to deviate from being linearly dependent on the applied field, by a factor of $\sqrt{T/T_e}$. Finally at sufficiently high fields, carriers start to interact with optical phonons and Eq. 78 is no longer accurate. The drift velocities for Ge and Si become less and less dependent on the applied field and approach a saturation velocity

$$v_s = \sqrt{\frac{8E_p}{3\pi m_0}} \approx 10^7 \text{ cm/s} \tag{80}$$

where E_p is the optical-phonon energy (listed in Appendix G).

To eliminate the discontinuity between the regimes covered by Eqs. 78–80, a single empirical formula is often used to describe the whole range, from low-field drift velocity to velocity saturation:[46]

$$v_d = \frac{\mu_0 \mathscr{E}}{[1 + (\mu_0 \mathscr{E}/v_s)^{C_2}]^{1/C_2}}. \tag{81}$$

The constant C_2 has a value near two for electrons and one for holes, and it is a function of temperature.

The velocity-field relationship is more complicated for GaAs, and we must consider its band structure (Fig. 4). A high-mobility valley ($\mu \approx$ 4,000 to 8,000 cm^2/V-s) is located at the Brillouin zone center, and a low-mobility satellite valley ($\mu \approx$

100 cm^2/V-s) along the $\langle 111 \rangle$-axes,[47] about 0.3 eV higher in energy. The difference in mobility is due to the different electron effective masses (Eq. 52): $0.063m_0$ in the lower valley and about $0.55m_0$ in the upper valley. As the field increases, the electrons in the lower valley can be excited to the normally unoccupied upper valley, resulting in a differential negative resistance in GaAs. The intervalley transfer mechanism, called transferred-electron effect, and the velocity-field relationship are considered in more detail in Chapter 10.

Figure 20a shows the measured room-temperature drift velocities versus electric field for high-purity (low impurity concentration) Si and GaAs. For high-level impurity dopings, the drift velocity or mobility at low fields is decreased due to impurity scattering. However, the velocity at high fields is essentially independent of impurity dopings, and it approaches a saturation value.[52] For Si the saturation velocities v_s for electrons and holes are about 1×10^7 cm/s. For GaAs a wide range of negative differential mobility exists for fields above 3×10^3 V/cm, and the high-field saturation velocity approaches 6×10^6 cm/s. Figure 20b shows the temperature dependence of electron saturation velocity. As the temperature increases, the saturation velocities for both Si and GaAs decrease.

Up to now, the drift velocities discussed are for steady-state condition where carriers go through enough scattering events to get to their equilibrium values. In modern devices, the critical dimension where carriers transit across becomes smaller and smaller. When this dimension becomes comparable to or shorter than the mean free path, *ballistic transport* is said to occur before carriers start to be scattered. Figure 21 shows the drift velocity as a function of distance. Without scattering, the velocity increases with time (and distance) according to $\approx q\mathscr{E}t/m^*$. At high fields, drift velocity can attain a higher value momentarily than that at steady state, within a short space (of the order of mean free path) and time (of the order of mean free time). This phenomenon is called velocity overshoot. (In literature, confusion might arise when the peak velocity of GaAs shown in Fig. 20a—the transferred-electron effect, is also called velocity overshoot.) At low fields, the acceleration of velocity is lower and when scattering starts to occur, the attained velocity is not that high and so velocity overshoot does not occur. Note that this shape of velocity overshoot is similar to that in the transferred-electron effect but the abscissa here is distance (or time) while that in the latter is electric field.

We next consider impact ionization. When the electric field in a semiconductor is increased above a certain value, the carriers gain enough energy to excite electron-hole pairs by a process called impact ionization. The threshold energy obviously has to be larger than the bandgap. This multiplication process is characterized by an ionization rate α defined as the number of electron-hole pairs generated by a carrier per unit distance traveled (Fig. 22). So for a primary carrier of electron traveling with a velocity v_n,

$$\alpha_n = \frac{1}{n}\frac{dn}{d(tv_n)} = \frac{1}{nv_n}\frac{dn}{dt}. \tag{82}$$

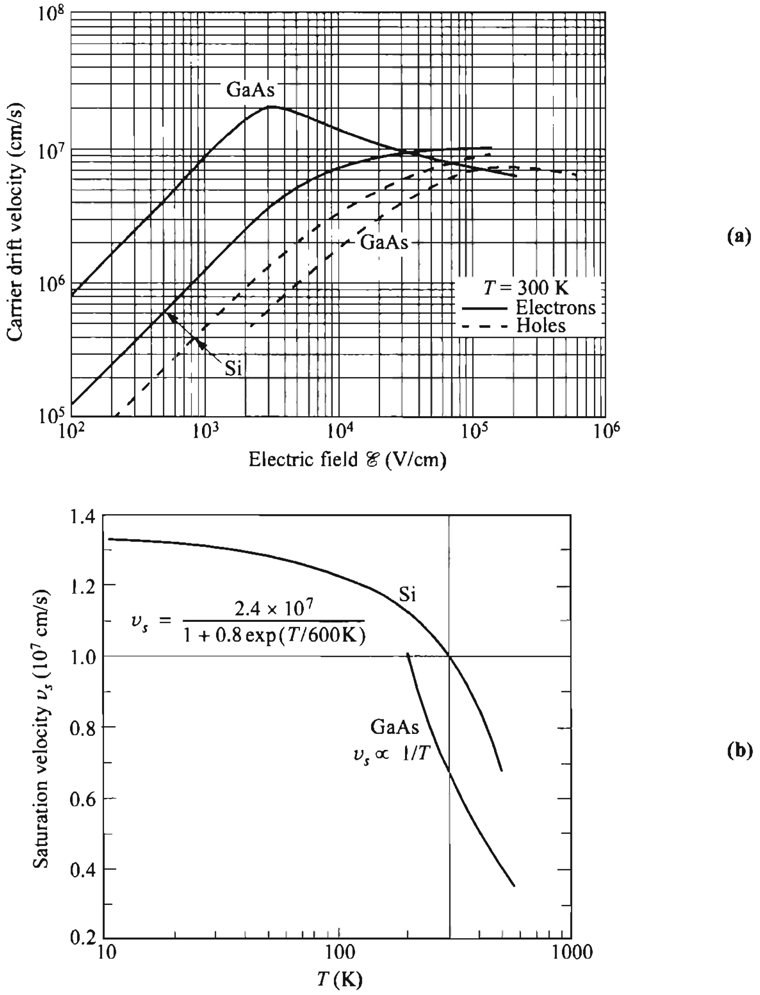


Fig. 20 (a) Measured carrier velocity versus electric field for high-purity Si and GaAs. For highly doped samples, the low-field velocities (mobilities) are lower than indicated here. In the high-field region, however, the velocity is essentially independent of dopings. (After Refs. 41, 48, 49, and 50.) (b) Saturated electron velocity versus temperature in Si and GaAs. (After Refs. 41 and 51.)

Considering both electrons and holes, the generation rate at any fixed location is given by

$$\begin{aligned} \frac{dn}{dt} = \frac{dp}{dt} &= \alpha_n n v_n + \alpha_p p v_p \\ &= \frac{\alpha_n J_n}{q} + \frac{\alpha_p J_p}{q} \quad . \end{aligned} \tag{83}$$

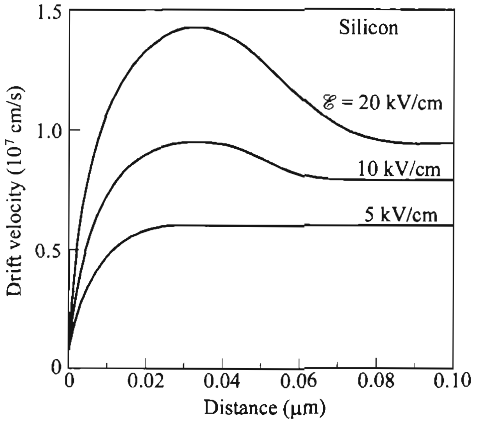


Fig. 21 Velocity overshoot in ultrashort distance. Similar behavior can be observed when the abscissa of distance is replaced with time. Example is for silicon. (After Ref. 53.)

Conversely, at any given time, the carrier density or current varies with distance and can be shown to be:

$$\frac{dJ_n}{dx} = \alpha_n J_n + \alpha_p J_p, \tag{84a}$$

$$\frac{dJ_p}{dx} = -\alpha_n J_n - \alpha_p J_p. \tag{84b}$$

The total current ($J_n + J_p$) remains constant over distance and $dJ_n/dx = -dJ_p/dx$.

The ionization rates α_n and α_p are strongly dependent on the electric field. A physical expression for the ionization rate is given by[54]

$$\alpha(\mathscr{E}) = \frac{q\mathscr{E}}{E_I}\exp\left\{-\frac{\mathscr{E}_I}{\mathscr{E}[1+(\mathscr{E}/\mathscr{E}_p)]+\mathscr{E}_T}\right\} \tag{85}$$

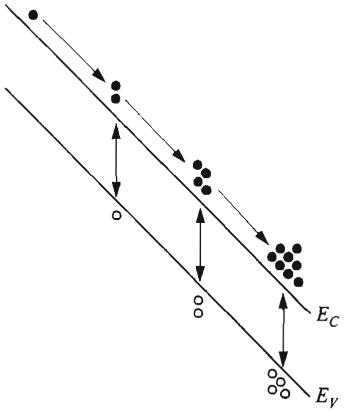


Fig. 22 Multiplication of electrons and holes from impact ionization, due to electrons (α_n) in this example ($\alpha_p = 0$).

where E_I is the high-field effective ionization threshold energy, and $\mathscr{E}_T$, $\mathscr{E}_p$, and $\mathscr{E}_I$ are threshold fields for carriers to overcome the decelerating effects of thermal, optical-phonon, and ionization scattering, respectively. For Si, the value of E_I is found to be 3.6 eV for electrons and 5.0 eV for holes. Over a limited field range, Equation 85 can be reduced to

$$\alpha(\mathscr{E}) = \frac{q\mathscr{E}}{E_I}\exp\left(-\frac{\mathscr{E}_I}{\mathscr{E}}\right), \qquad \text{if} \qquad \mathscr{E}_p > \mathscr{E} > \mathscr{E}_T , \tag{86}$$

or

$$\alpha(\mathscr{E}) = \frac{q\mathscr{E}}{E_I}\exp\left(-\frac{\mathscr{E}_I\mathscr{E}_p}{\mathscr{E}^2}\right), \qquad \text{if } \mathscr{E} > \mathscr{E}_p \text{ and } \mathscr{E} > \sqrt{\mathscr{E}_p\mathscr{E}_T}. \tag{87}$$

Figure 23a shows the experimental results of the ionization rates for Ge, Si, SiC, and GaN. Figure 23b shows the measured ionization rates of GaAs and a few other binary and ternary compounds. These results are obtained by using photomultiplication measurements on *p-n* junctions. Note that for certain semiconductors, such as GaAs, the ionization rate is a function of crystal orientation. There is also a general trend that the ionization rate decreases with increasing bandgap. It is for this reason that materials of higher bandgaps generally yield higher breakdown voltage. Note that Eq. 86 is applicable to most semiconductors shown in Fig. 23, except GaAs and GaP, for which Eq. 87 is applicable.

At a given electric field, the ionization rate decreases with increasing temperature. Figure 24 shows the theoretical predicted electron ionization rates in silicon as an example, together with the experimental results at three different temperatures.

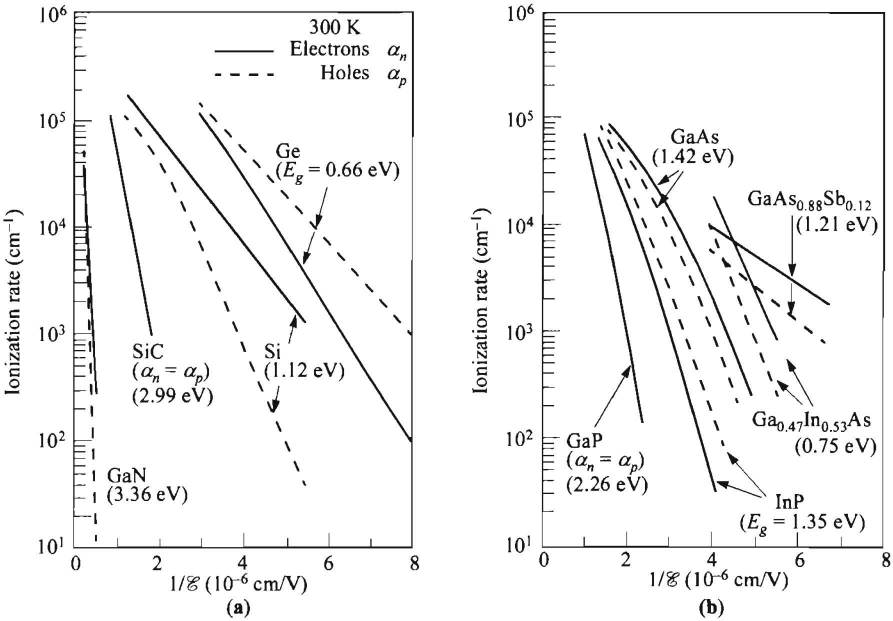
1.5.4 Recombination, Generation, and Carrier Lifetimes

Whenever the thermal-equilibrium condition of a semiconductor system is disturbed (i.e., $pn \neq n_i^2$), processes exist to restore the system to equilibrium (i.e., $pn = n_i^2$). These processes are recombination when $pn > n_i^2$ and thermal generation when $pn < n_i^2$. Figure 25a illustrates the band-to-band electron-hole recombination. The energy of an electron in transition from the conduction band to the valence band is conserved by emission of a photon (radiative process) or by transfer of the energy to another free electron or hole (Auger process). The former process is the inverse of direct optical absorption, and the latter is the inverse of impact ionization.

Band-to-band transitions are more probable for direct-bandgap semiconductors which are more common among III-V compounds. For this type of transition, the recombination rate is proportional to the product of electron and hole concentrations, given by

$$R_e = R_{ec}pn. \tag{88}$$

The term R_{ec}, called the *recombination coefficient*, is related to the thermal generation rate G_{th} by



(a) **(b)**

Fig. 23 Ionization rates at 300 K versus reciprocal electric field for Si, GaAs, and some IV-IV and III-V compound semiconductors. (After Refs. 55–65.)

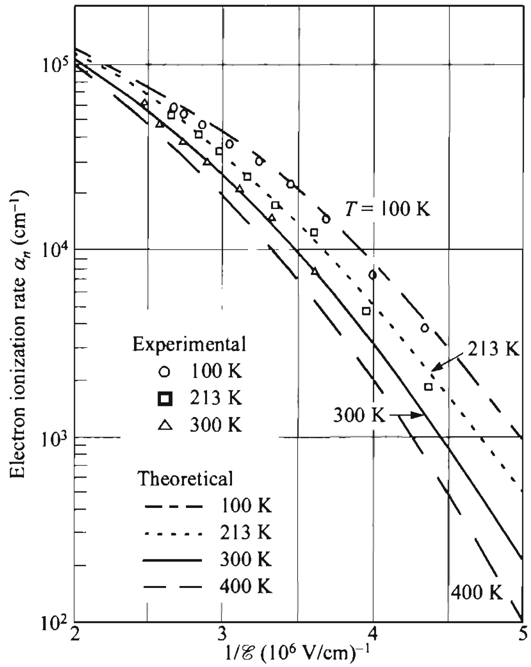


Fig. 24 Electron ionization rate versus reciprocal electric field in Si for four temperatures. (After Ref. 66.)

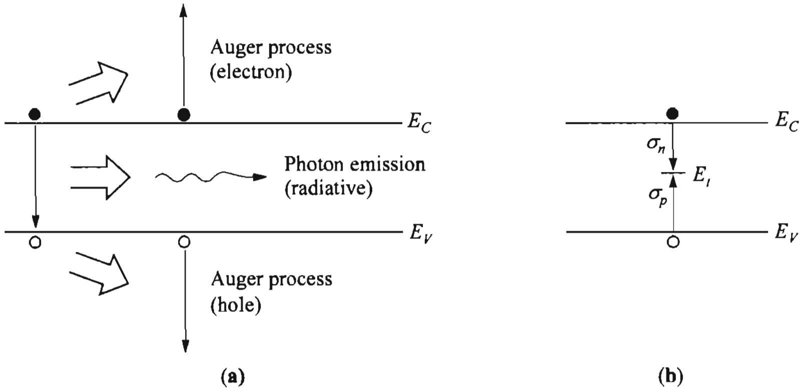


Fig. 25 Recombination processes (the reverse are generation processes). (a) Band-to-band recombination. Energy is exchanged to a radiative or Auger process. (b) Recombination through single-level traps (nonradiative).

$$R_{ec} = \frac{G_{th}}{n_i^2}. \tag{89}$$

R_{ec} is a function of temperature and is also dependent on the band structure of the semiconductor. A direct-bandgap semiconductor, being more efficient in band-to-band transitions, has a much larger R_{ec} ($\approx 10^{-10}$ cm^3/s) than an indirect-bandgap semiconductor ($\approx 10^{-15}$ cm^3/s). In thermal equilibrium, since $pn = n_i^2$, $R_e = G_{th}$ and the net transition rate $U (= R_e - G_{th})$ equals zero. Under low-level injection, defined as the case where the excess carriers $\Delta p = \Delta n$ are fewer than the majority carriers, for an n-type material $p_n = p_{no} + \Delta p$ and $n_n \approx N_D$, the net transition rate is given by

$$\begin{aligned} U &= R_e - G_{th} = R_{ec}(pn - n_i^2) \\ &\approx R_{ec}\Delta p N_D \equiv \frac{\Delta p}{\tau_p} \end{aligned} \tag{90}$$

where the carrier lifetime for holes

$$\tau_p = \frac{1}{R_{ec}N_D}, \tag{91a}$$

and in p-type material,

$$\tau_n = \frac{1}{R_{ec}N_A}. \tag{91b}$$

However, in indirect-bandgap semiconductors such as Si and Ge, the dominant transitions are indirect recombination/generation via bulk traps, of density N_t and energy E_t present within the bandgap (Fig. 25b). The single-level recombination can be described by two processes—electron capture and hole capture. The net transition rate can be described by the Shockley-Read-Hall statistics[67–69] as

$$U = \frac{\sigma_n \sigma_p v_{th} N_t (pn - n_i^2)}{\sigma_n \left[n + n_i \exp\left(\frac{E_t - E_i}{kT}\right)\right] + \sigma_p \left[p + n_i \exp\left(\frac{E_i - E_t}{kT}\right)\right]} \tag{92}$$

where σ_n and σ_p are the electron and hole capture cross sections, respectively. Without deriving this equation, some qualitative observations can be made on the final form. First, the net transition rate is proportional to $pn - n_i^2$, similar to Eq. 90, and the sign determines whether there is net recombination or generation. Second, U is maximized when $E_t = E_i$, indicating for an energy spectrum of bulk traps, only those near the mid-gap are effective recombination/generation centers. Considering only these traps, Eq. 92 is reduced to

$$U = \frac{\sigma_n \sigma_p v_{th} N_t (pn - n_i^2)}{\sigma_n (n + n_i) + \sigma_p (p + n_i)} . \tag{93}$$

Again for low-level injection in n-type semiconductors, the net recombination rate becomes

$$\begin{aligned} U &= \frac{\sigma_n \sigma_p v_{th} N_t [(p_{no} + \Delta p) n - n_i^2]}{\sigma_n n} \\ &\approx \sigma_p v_{th} N_t \Delta p \equiv \frac{\Delta p}{\tau_p} \end{aligned} \tag{94}$$

where

$$\tau_p = \frac{1}{\sigma_p v_{th} N_t} . \tag{95a}$$

Similarly for a p-type semiconductor, the electron lifetime is given by

$$\tau_n = \frac{1}{\sigma_n v_{th} N_t} . \tag{95b}$$

As expected, the lifetime arising from indirect transitions is inversely proportional to the trap density N_t, while in the previous case, the lifetime from direct transitions is inversely proportional to the doping concentration (Eqs. 91a and 91b).

For multiple-level traps, the recombination processes have gross qualitative features that are similar to those of the single-level case. However, the behavioral details are different, particularly in the high-level injection condition (i.e., where $\Delta n = \Delta p$ approaches the majority-carrier concentration), where the asymptotic lifetime is an average of the lifetimes associated with all the positively charged, negatively charged, and neutral trapping levels.

For high-level injection ($\Delta n = \Delta p > n$ and p), the carrier lifetime for band-to-band recombination becomes

$$\tau_n = \tau_p = \frac{1}{R_{ec} \Delta n} . \tag{96}$$

The lifetime resulting from traps can be derived from Eq. 93 to be

$$\tau_n = \tau_p = \frac{\sigma_n + \sigma_p}{\sigma_n \sigma_p v_{th} N_t}. \tag{97}$$

Comparing Eq. 97 to Eqs. 95a and 95b, the lifetime is actually higher with high-level injection. It is interesting to note that the lifetime due to band-to-band recombination decreases with injection level, while that due to trap recombination increases with injection level.

Equations 95a and 95b have been verified experimentally by using solid-state diffusion and high-energy radiation. Many impurities have energy levels close to the middle of the bandgap (Fig. 10). These impurities are efficient recombination centers. A typical example is gold in silicon;[70] the minority-carrier lifetime decreases linearly with the gold concentration over the range of 10^{14} to 10^{17} cm^{-3}, where τ decreases from about 2×10^{-6} s to 2×10^{-9} s. This effect is sometimes advantageous, as in some high-speed applications when a short lifetime to reduce the charge storage time is a desirable feature. Another method of shortening the minority-carrier lifetime is high-energy-particle irradiation, which causes displacement of host atoms and damage to the lattice. These, in turn, introduce energy levels in the bandgap. For example, electron irradiation in Si gives rise to an acceptor level at 0.4 eV above the valence band and a donor level at 0.36 eV below the conduction band. Also neutron irradiation creates an acceptor level at 0.56 eV; and deuteron irradiation gives rise to an interstitial state with an energy level 0.25 eV above the valence band. Similar results are obtained for Ge, GaAs, and other semiconductors. Unlike the solid-state diffusion, the radiation-induced trapping centers may be annealed out at relatively low temperatures.

When carriers are below their thermal-equilibrium values, i.e., $pn < n_i^2$, generation of carriers rather than recombination of excess carriers will occur. The generation rate can be found by starting with Eq. 93,

$$U = -\frac{\sigma_p \sigma_n v_{th} N_t n_i}{\sigma_p[1 + (p/n_i)] + \sigma_n[1 + (n/n_i)]} \equiv -\frac{n_i}{\tau_g} \tag{98}$$

where the generation carrier lifetime τ_g is equal to

$$\begin{aligned} \tau_g &= \frac{1 + (n/n_i)}{\sigma_p v_{th} N_t} + \frac{1 + (p/n_i)}{\sigma_n v_{th} N_t} \\ &= \left(1 + \frac{n}{n_i}\right)\tau_p + \left(1 + \frac{p}{n_i}\right)\tau_n. \end{aligned} \tag{99}$$

Depending on the electron and hole concentrations, the generation lifetime can be much longer than the recombination lifetime and has a minimum value of roughly twice that of the recombination lifetime, when both n and p are much smaller than n_i.

The minority-carrier lifetime τ has generally been measured using the photoconductive (PC) effect[71] or the photoelectromagnetic (PEM) effect[72]. The basic equation for the PC effect is given by

$$J_{\mathrm{PC}} = q(\mu_n + \mu_p)\Delta n \mathscr{E} = q(\mu_n + \mu_p)\frac{G_e}{\tau}\mathscr{E} \tag{100}$$

where J_{PC} is the incremental current density as a result of illumination with generation rate G_e, and $\mathscr{E}$ is the applied electric field along the sample. The quantity Δn is the incremental carrier density or the number of electron-hole pairs per volume created by the illumination, which equals the product of the generation rate G_e and the lifetime τ, or $\Delta n = \tau G_e$. For the PEM effect we measure the short-circuit current, which appears when a constant magnetic field $\mathscr{B}_z$ is applied perpendicular to the direction of incoming radiation. The current density is given by

$$J_{\mathrm{PEM}} = q(\mu_n + \mu_p)\mathscr{B}_z\frac{D}{L_d}\tau G_e = q(\mu_n + \mu_p)\mathscr{B}_z\sqrt{D\tau}\,G_e \tag{101}$$

where D and L_d [$\equiv (D\tau)^{1/2}$] are the diffusion coefficient and the diffusion length, to be discussed in the next section. Another approach to measure the carrier lifetime will be discussed in Section 1.8.2.

1.5.5 Diffusion

In the preceding section the excess carriers are uniform in space. In this section, we discuss the situations where excess carriers are introduced locally, causing a condition of nonuniform carriers. Examples are local injection of carriers from a junction, and nonuniform illumination. Whenever there exists a gradient of carrier concentration, a process of diffusion occurs by which the carriers migrate from the region of high concentration toward the region of low concentration, to drive the system toward a state of uniformity. This flow or flux of carriers, taking electrons as an example, is governed by the Fick's law,

$$\left.\frac{d\Delta n}{dt}\right|_x = -D_n\frac{d\Delta n}{dx}, \tag{102}$$

and is proportional to the concentration gradient. The proportionality constant is called the diffusion coefficient or diffusivity D_n. This flux of carriers constitutes a diffusion current, given by

$$J_n = qD_n\frac{d\Delta n}{dx}, \tag{103a}$$

and

$$J_p = -qD_p\frac{d\Delta p}{dx}. \tag{103b}$$

Physically, diffusion is due to random thermal motion of carriers as well as scattering. Because of this, we have

$$D = v_{th}\tau_m. \tag{104}$$

One also expects certain relationship between the diffusion coefficient and mobility. To derive such a relationship, we consider an n-type semiconductor with nonuniform doping concentration but without an external applied field. The zero net current necessitates that the drift current exactly balances the diffusion current,

$$qn\mu_n\mathscr{E} = -qD_n\frac{dn}{dx}. \tag{105}$$

In this case, the electric field is created by the nonuniform doping ($\mathscr{E} = dE_C/qdx$, and E_F is constant for equilibrium). Using Eq. 21 for n, we obtain

$$\begin{aligned} \frac{dn}{dx} &= \frac{-q\mathscr{E}}{kT}N_C\exp\left(-\frac{E_C-E_F}{kT}\right) \\ &= \frac{-q\mathscr{E}}{kT}n . \end{aligned} \tag{106}$$

Substituting this into Eq. 105 will give the relationship

$$D_n = \left(\frac{kT}{q}\right)\mu_n. \tag{107a}$$

Similarly for p-type semiconductor, one can derive

$$D_p = \left(\frac{kT}{q}\right)\mu_p. \tag{107b}$$

These are known as the Einstein relation (valid for nondegenerate semiconductors). At 300 K, $kT/q = 0.0259$ V, and values of D are readily obtainable from the mobility results shown in Fig. 15.

Another parameter closely related to diffusion is the diffusion length,

$$L_d = \sqrt{D\tau} . \tag{108}$$

In common diffusion problems arising from some fixed injection source as a boundary condition, the resultant concentration profile is exponential in nature with distance, with a characteristic length of L_d. This diffusion length can also be viewed as the distance carriers can diffuse in a carrier lifetime before they are annihilated.

1.5.6 Thermionic Emission

Another current conduction mechanism is *thermionic emission*. It is a majority-carrier current and is always associated with a potential barrier. Note that the critical parameter is the barrier height, not the shape of the barrier. The most-common device is the Schottky-barrier diode or metal-semiconductor junction (see Chapter 3). Referring to Fig. 26, for the thermionic emission to be the controlling mechanism, the criterion is that collision or the drift-diffusion process within the barrier layer to be negligible. Equivalently, the barrier width has to be narrower than the mean free path, or in the case of a triangular barrier, the slope of the barrier be reasonably steep such that a drop in kT in energy is within the mean free path. In addition, after the carriers are injected over the barrier, the diffusion current in that region must not be the lim-

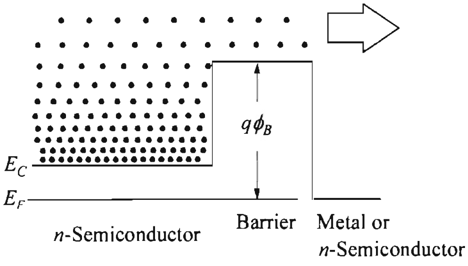


Fig. 26 Energy-band diagram showing thermionic emission of electrons over the barrier. Note that the shape of the barrier (shown as rectangular) does not matter.

iting factor. Therefore, the region behind the barrier must be another *n*-type semiconductor or a metal layer.

Due to Fermi-Dirac statistics, the density of electrons (for *n*-type substrate) decreases exponentially as a function of their energy above the conduction band edge. At any finite (nonzero) temperature, the carrier density at any finite energy is not zero. Of special interest here is the integrated number of carriers above the barrier height. This portion of the thermally generated carriers are no longer confined by the barrier so they contribute to the thermionic-emission current. The total electron current over the barrier is given by (see Chapter 3)

$$J = A^* T^2 \exp\left(-\frac{q\phi_B}{kT}\right). \tag{109}$$

where ϕ_B is the barrier height, and

$$A^* \equiv \frac{4\pi q m^* k^2}{h^3} \tag{110}$$

is called the effective Richardson constant and is a function of the effective mass. The A^* can be further modified by quantum-mechanical tunneling and reflection.

1.5.7 Tunneling

Tunneling is a quantum-mechanical phenomenon. In classical mechanics, carriers are completely confined by the potential walls. Only those carriers with excess energy higher than the barriers can escape, as in the case of thermionic emission discussed above. In quantum mechanics, an electron can be represented by its wavefunction. The wavefunction does not terminate abruptly on a wall of finite potential height and it can penetrate into and through the barrier (Fig. 27). The probability of electron tunneling through a barrier of finite height and width is thus not zero.

To calculate the tunneling probability, the wavefunction ψ has to be determined from the Schrödinger equation

$$\frac{d^2\psi}{dx^2} + \frac{2m^*}{\hbar^2}[E - U(x)]\psi = 0. \tag{111}$$

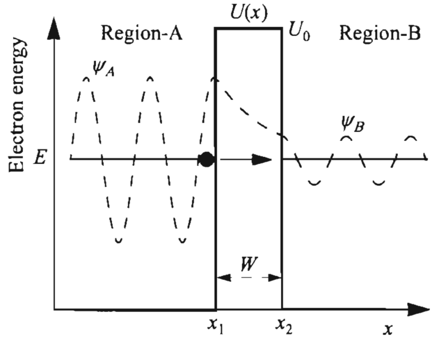


Fig. 27 Wavefunctions showing electron tunneling through a rectangular barrier.

In the case of a simple rectangular barrier of height U_0 and width W, ψ has a general form of $\exp(\pm ikx)$ where $k = \sqrt{2m^*(E-U_0)}/\hbar$. Note that for tunneling, the energy E is below the barrier U_0 so that the term within the square root is negative and k is imaginary. The solution of the wavefunctions and the tunneling probability are calculated to be

$$T_t = \frac{|\psi_B|^2}{|\psi_A|^2} = \left[1 + \frac{U_0^2 \sinh^2(|k|W)}{4E(U_0-E)}\right]^{-1}$$
$$\approx \frac{16E(U_0-E)}{U_0^2}\exp\left(-2\sqrt{\frac{2m^*(U_0-E)}{\hbar^2}}\,W\right). \tag{112}$$

For more complicated barrier shapes, simplification of the Schrödinger equation is made by the WKB (Wentzel-Kramers-Brillouin) approximation if the potential $U(x)$ does not vary rapidly. The wavefunction now has a general form of $\exp\int ik(x)dx$. The tunneling probability can be calculated by

$$T_t = \frac{|\psi_B|^2}{|\psi_A|^2} \approx \exp\left\{-2\int_{x_1}^{x_2} |k(x)|\,dx\right\}$$
$$\approx \exp\left\{-2\int_{x_1}^{x_2} \sqrt{\frac{2m^*}{\hbar^2}[U(x)-E]}\,dx\right\}. \tag{113}$$

Together with known tunneling probability, the tunneling current J_t can be calculated from the product of the number of available carriers in the originating Region-A (Fig. 27), and the number of empty states in the destination Region-B,

$$J_t = \frac{qm^*}{2\pi^2\hbar^3}\int F_A N_A T_t(1-F_B)N_B\,dE \tag{114}$$

where F_A, F_B, N_A, and N_B represent the Fermi-Dirac distributions and densities of states in the corresponding regions.

1.5.8 Space-Charge Effect

The space charge in a semiconductor is determined by both the doping concentrations and the free-carrier concentrations,

$$\rho = (p - n + N_D - N_A)q \quad . \tag{115}$$

In the neutral region of a semiconductor, $n = N_D$ and $p = N_A$, so that the space-charge density is zero. In the vicinity of a junction formed by different materials, dopant types, or doping concentrations, n and p could be smaller or larger than N_D and N_A, respectively. In the depletion approximation, n and p are assumed zero so that the space charge is equal to the majority-carrier doping level. Under bias, the carrier concentrations n and p can be increased beyond their values in equilibrium. When the injected n or p is larger than its equilibrium value as well as the doping concentration, the *space-charge effect* is said to occur. The injected carriers thus control the space charge and the electric-field profile. This results in a feedback mechanism where the field drives the current, which in turn sets up the field. The space-charge effect is more common in lightly doped materials, and it can occur outside the depletion region.

In the presence of a space-charge effect, if the current is dominated by the drift component of the injected carriers, it is called the *space-charge-limited current*. Since it is a drift current, it is given by, in the case of electron injection,

$$J = qnv . \tag{116}$$

The space charge again is determined by the injected carriers giving rise to the Poisson equation of the form

$$\frac{d^2\psi_i}{dx^2} = \frac{qn}{\varepsilon_s} \quad . \tag{117}$$

The carrier velocity v is related to the electric field by different functions, depending on the field strength. In the low-field mobility regime,

$$v = \mu\mathscr{E} \quad . \tag{118}$$

In the velocity-saturation regime, velocity v_s is independent of the field. In the limit of ultra-short sample or time scale, we have the ballistic regime where there is no scattering, and

$$v = \sqrt{\frac{2qV}{m^*}} \quad . \tag{119}$$

From Eqs. 116–119, the space-charge-limited current in the mobility regime (the Mott-Gurney law) can be solved to be (see Vol. 4 of Ref. 4)

$$J = \frac{9\varepsilon_s\mu V^2}{8L^3} , \tag{120}$$

in the velocity-saturation regime

$$J = \frac{2\varepsilon_s v_s V}{L^2} , \tag{121}$$

and in the ballistic regime (the Child-Langmuir law)

$$J = \frac{4\varepsilon_s}{9L^2}\left(\frac{2q}{m^*}\right)^{1/2} V^{3/2}. \tag{122}$$

Here L is the length of the sample in the direction of the current flow. Note that the voltage dependence is different in these regimes.

1.6 PHONON, OPTICAL, AND THERMAL PROPERTIES

In the preceding section we considered different carrier transport mechanisms in semiconductors. In this section we briefly consider other effects and properties of semiconductors that are important to the operation of semiconductor devices.

1.6.1 Phonon Spectra

Phonons are quanta of lattice vibrations, mainly resulting from the lattice thermal energy. Similar to photons and electrons, they have characteristic frequencies (or energy) and wave numbers (momentum or wavelengths). It is known that, as a demonstration in a one-dimensional lattice, with only nearest-neighbor coupling and two different masses m_1 and m_2 placed alternately, the frequencies of oscillation are given by[3]

$$\nu_\pm = \sqrt{\alpha_f}\left[\left(\frac{1}{m_1}+\frac{1}{m_2}\right) \pm \sqrt{\left(\frac{1}{m_1}+\frac{1}{m_2}\right)^2 - \frac{4\sin^2(k_{ph}a/2)}{m_1 m_2}}\right]^{1/2} \tag{123}$$

where α_f is the force constant of the Hooke's law, k_{ph} the phonon wave number, and a the lattice spacing. The frequency ν_- is proportional to k_{ph} near $k_{ph} = 0$. This branch is the acoustic branch, because it is the long-wavelength vibration of the lattice and the velocity ω/k is near that of sound in such a medium. The frequency ν_+ tends to be a constant $\approx [2\alpha_f(1/m_1 + 1/m_2)]^{1/2}$ as k_{ph} approaches zero. This branch, separated considerably from the acoustic mode, is the optical branch, because the frequency ν_+ is generally in the optical range. For the acoustic mode the two sublattices of the atoms with different masses move in the same direction, whereas for the optical mode they move in opposite directions.

The total number of acoustic modes is equal to the dimension times the number of atoms per cell. For a realistic three-dimensional lattice with one atom per primitive cell, such as a simple cubic, body-centered, or face-centered cubic lattice, only three acoustic modes exist. For a three-dimensional lattice with two atoms per primitive cell, such as Si and GaAs, three acoustic modes and three optical modes exist. Longitudinally polarized modes are modes with the displacement vectors of each atom along the direction of the wave vector; thus we have one longitudinal acoustic mode (LA) and one longitudinal optical mode (LO). Modes with atoms moving in the planes normal to the wave vector are called transversely polarized modes. We have two transverse acoustic modes (TA) and two transverse optical modes (TO).

Figure 28 shows the measured results for Si and GaAs in one of the crystal directions. The range of $k_{ph} = \pm\pi/a$ defines the Brillouin zone outside which the frequency-k_{ph} relationship repeats itself. Note that at small values of k_{ph}, for both LA and TA modes, the energies (or frequencies) are proportional to k_{ph}. The longitudinal optical phonon energy at $k_{ph} = 0$ is the first-order Raman scattering energy. Their values are 0.063 eV for Si and 0.035 eV for GaAs. Appendix G lists these results, together with other important properties.

1.6.2 Optical Properties

Optical measurement constitutes the most-important means of determining the band structures of semiconductors. Photon-induced electronic transitions can occur between different bands, which lead to the determination of the energy bandgap, or within a single band such as the free-carrier absorption. Optical measurements can also be used to study lattice vibrations (phonons). The optical properties of semiconductor are characterized by the complex refractive index,

$$\bar{n} = n_r - ik_e . \qquad (124)$$

The real part of the refractive index n_r determines the propagation velocity (v and wavelength λ) in the medium (assuming ambient is a vacuum having wavelength λ_0)

$$n_r = \frac{c}{v} = \frac{\lambda_0}{\lambda} . \qquad (125)$$

The imaginary part k_e, called the extinction coefficient, determines the absorption coefficient

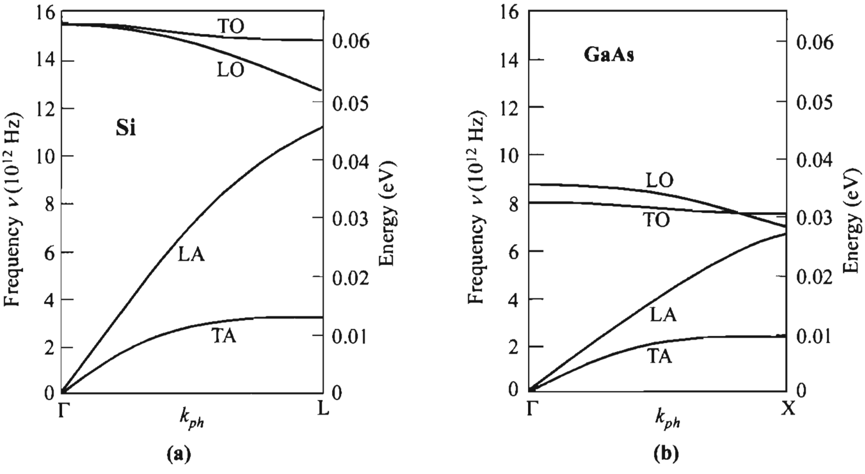


Fig. 28 Measured phonon spectra in (a) Si (After Ref. 73.) and (b) GaAs (After Ref. 74.). TO and LO stand for transverse and longitudinal optical modes, and TA and LA for transverse and longitudinal acoustic modes.

$$\alpha = \frac{4\pi k_e}{\lambda}. \tag{126}$$

In semiconductors, the absorption coefficient is a strong function of the wavelength or photon energy. Near the absorption edge, the absorption coefficient can be expressed as[5]

$$\alpha \propto (h\nu - E_g)^\gamma \tag{127}$$

where $h\nu$ is the photon energy and γ is a constant. There exist two types of band-to-band transitions: allowed and forbidden. (Forbidden transitions take into account the small but finite momentum of photons and are much less probable.) For direct-bandgap materials, transitions mostly occur between two bands of the same k value, as transitions (a) and (b) in Fig. 29. While allowed direct transitions can occur in all k values, forbidden direct transitions can only occur at $k \neq 0$. In the one-electron approximation, γ equals 1/2 and 3/2 for allowed and forbidden direct transitions, respectively. Note that for $k = 0$ at which the bandgap is defined, only allowed transition ($\gamma = 1/2$) occurs and thus it is used in determining the bandgap experimentally. For indirect transitions [transition (c) in Fig. 29], phonons are involved in order to conserve momentum. In these transitions, phonons (with energy E_p) are either absorbed or emitted, and the absorption coefficient is modified to

$$\alpha \propto (h\nu - E_g \pm E_p)^\gamma . \tag{128}$$

Here the constant γ equals 2 and 3 for allowed and forbidden indirect transitions, respectively.

In addition, increased absorption peaks and steps can be due to formation of excitons, which are bound electron-hole pairs with energy levels within the bands that move through the crystal lattice as a unit. Near the absorption edge, where the values of $(E_g - h\nu)$ become comparable with the binding energy of an exciton, the Coulomb interaction between the free electron and hole must be taken into account. The photon energy required for absorption is lowered by this binding energy. For $h\nu \lessapprox E_g$ the

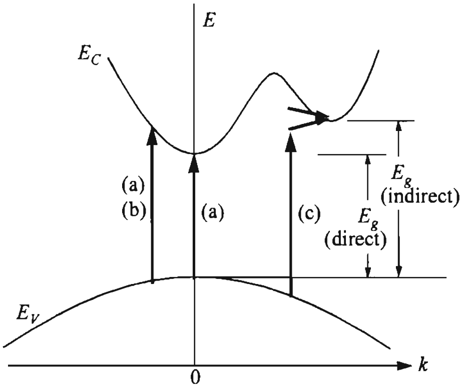


Fig. 29 Optical transitions: (a) allowed and (b) forbidden direct transitions; (c) indirect transition involving phonon emission (upper arrow) and phonon absorption (lower arrow).

absorption merges continuously into the fundamental absorption. When $h\nu \gg E_g$, higher energy bands participate in the transition processes, and complicated band structures are reflected in the absorption coefficient.

Figure 30 plots the experimental absorption coefficients α near and above the fundamental absorption edge (band-to-band transition) for Si and GaAs. The shift of the curves toward higher photon energies at lower temperature is associated with the temperature dependence of the bandgap (Fig. 6). An α of 10^4 cm^{-1} means that 63% of light will be absorbed in one micron of semiconductor.

When light passes through a semiconductor, absorption of light and generation of electron-hole pairs (G_e) occur, and the light intensity P_{op} diminishes with distance according to

$$\frac{dP_{op}(x)}{dx} = -\alpha P_{op}(x) = G_e h\nu. \tag{129}$$

Solution of the above gives an exponential decay of intensity

$$P_{op}(x) = P_0(1-R)\exp(-\alpha x) \tag{130}$$

where P_0 is the external incident light intensity and R is the reflection of the ambient-semiconductor interface at normal incidence,

$$R = \frac{(1-n_r)^2 + k_e^2}{(1+n_r)^2 + k_e^2}. \tag{131}$$

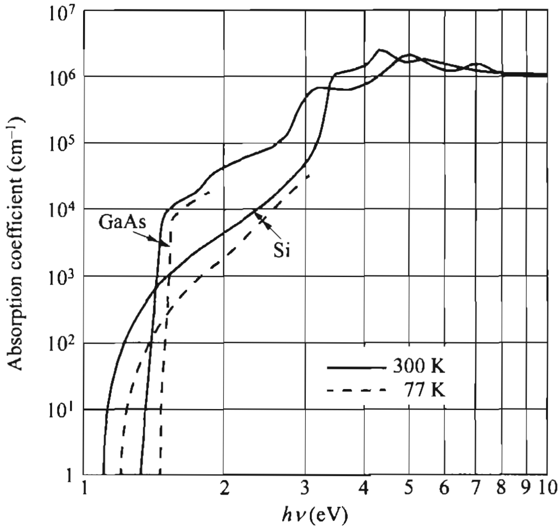


Fig. 30 Measured absorption coefficients near and above the fundamental absorption edge for Si and GaAs. (After Refs. 75–78.)

In a semiconductor sample of thickness W where the product αW is not large, multiple reflections will occur between the two interfaces. Summing up all the light components in the backward direction, the total reflection coefficient is calculated to be

$$R_\Sigma = R\left[1 + \frac{(1-R)^2\exp(-2\alpha W)}{1-R^2\exp(-2\alpha W)}\right], \tag{132}$$

and the total transmission coefficient is given by

$$T_\Sigma = \frac{(1-R)^2\exp(-\alpha W)}{1-R^2\exp(-2\alpha W)}. \tag{133}$$

The transmission coefficient T_Σ and the reflection coefficient R_Σ are two important quantities generally measured. By analyzing the T_Σ-λ or R_Σ-λ data at normal incidence, or by making observations on R_Σ or T_Σ for different angles of incidence, both n_r and k_e can be obtained and related to the transition energy between bands.

1.6.3 Thermal Properties

When a temperature gradient exists in a semiconductor in addition to an applied electric field, the total current density (in one dimension) is[5]

$$J = \sigma\left(\frac{1}{q}\frac{dE_F}{dx} - \mathscr{P}\frac{dT}{dx}\right) \tag{134}$$

where $\mathscr{P}$ is the thermoelectric power, so named to indicate that for an open-circuit condition the net current is zero and an electric field is generated by the temperature gradient. For a nondegenerate semiconductor with a mean free time between collisions $\tau_m \propto E^{-s}$ as discussed previously, the thermoelectric power is given by

$$\mathscr{P} = -\frac{k}{q}\left\{\frac{[\frac{5}{2} - s + \ln(N_C/n)]n\mu_n - [\frac{5}{2} - s - \ln(N_V/p)]p\mu_p}{n\mu_n + p\mu_p}\right\} \tag{135}$$

(k is Boltzmann constant). This equation indicates that the thermoelectric power is negative for n-type semiconductors and positive for p-type semiconductors, a fact often used to determine the conduction type of a semiconductor. The thermoelectric power can also be used to determine the resistivity and the position of the Fermi level relative to the band edges. At room temperature the thermoelectric power $\mathscr{P}$ of p-type silicon increases with resistivity: 1 mV/K for a 0.1 Ω-cm sample and 1.7 mV/K for a 100 Ω-cm sample. Similar results (except a change of the sign for $\mathscr{P}$) can be obtained for n-type silicon samples.

Another important thermal effect is thermal conduction. It is a diffusion type of process where the heat flow Q is driven by the temperature gradient

$$Q = -\kappa\frac{dT}{dx}. \tag{136}$$

The thermal conductivity κ has the major components of phonon (lattice) conduction κ_L and mixed free-carrier conduction κ_M of electrons and holes,

$$\kappa = \kappa_L + \kappa_M. \tag{137}$$

The lattice contribution is carried out by diffusion and scattering of phonons. These scattering events include many types, such as phonon-to-phonon, phonon-to-defects, phonon-to-carriers, boundaries and surfaces, and so on. The overall effect can be interpreted as

$$\kappa_L = \frac{1}{3}C_v v_{ph}\lambda_{ph} \tag{138}$$

where C_v is the specific heat, v_{ph} the phonon velocity, and λ_{ph} the phonon mean free path. The contribution due to mixed carriers, if $\tau_m \propto E^{-s}$ holds for both electron and hole scattering, is given by

$$\kappa_M = \frac{(\frac{5}{2}-s)k^2\sigma T}{q^2} + \frac{k^2\sigma T}{q^2}\frac{[5-2s+(E_g/kT)]^2 np\mu_n\mu_p}{(n\mu_n+p\mu_p)^2}. \tag{139}$$

Figure 31 shows the measured thermal conductivity as a function of lattice temperature for Si and GaAs. Appendix G lists the room-temperature values. The contributions of conduction carriers to the thermal conductivity are in general quite small,

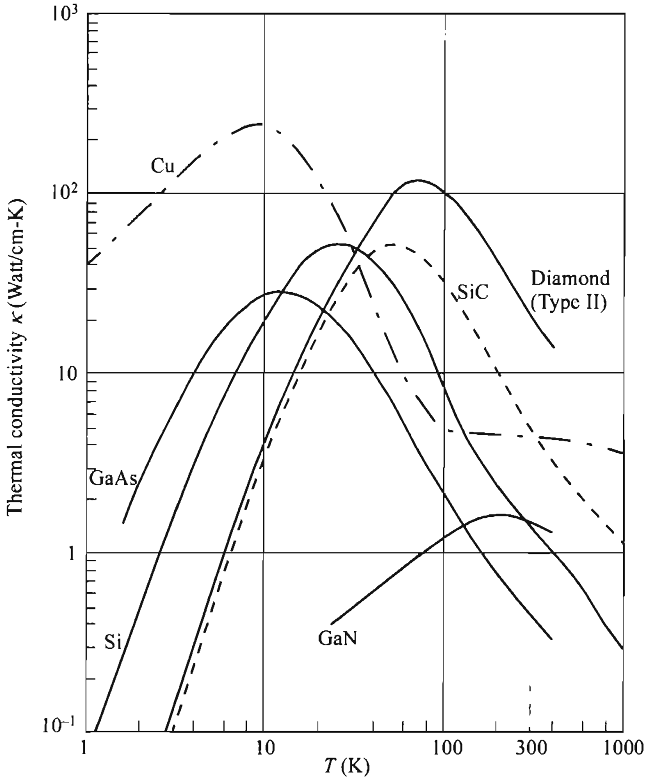


Fig. 31 Measured thermal conductivity versus temperature for pure Si, GaAs, SiC, GaN, Cu, and diamond (Type II). (After Refs. 79–83.)

so the general temperature dependence follows that of κ_L and has a inverted V-shape. At low temperatures, the specific heat has a T^3 dependence and κ goes up sharply. At high temperatures, phonon-phonon scattering dominates and λ_{ph} (and κ_L) drops according to $1/T$. Figure 31 also shows the thermal conductivities for Cu, diamond, SiC, and GaN. Copper is the most commonly used metal for thermal conduction in *p-n* junction devices; diamond has the highest room-temperature thermal conductivity known to date and is useful as the thermal sink for semiconductor lasers and IMPATT oscillators. SiC and GaN are important semiconductors for power devices.

1.7 HETEROJUNCTIONS AND NANOSTRUCTURES

A heterojunction is a junction formed between two dissimilar semiconductors. For semiconductor-device applications, the difference in energy gap provides another degree of freedom that produces many interesting phenomena. The successful applications of heterojunctions in various devices is due to the capability of epitaxy technology to grow lattice-matched semiconductor materials on top of one another with virtually no interface traps. Heterojunctions have been widely used in various device applications. The underlying physics of epitaxial heterojunction is matching of the lattice constants. This is a physical requirement in atom placement. Severe lattice mismatch will cause dislocations at the interface and results in electrical defects such as interface traps. The lattice constants of some common semiconductors are shown in Fig. 32, together with their energy gaps. A good combination for heterojunction devices is two materials of similar lattice constants but different E_g. As can be seen, GaAs/AlGaAs (or /AlAs) is a good example.

It turns out that if the lattice constants are not severely mismatched, good-quality heteroepitaxy can still be grown, provided that the epitaxial-layer thickness is small enough. The amount of lattice mismatch and the maximum allowed epitaxial layer are directly related. This can be explained with the help of Fig. 33. For a relaxed, thick heteroepitaxial layer, dislocations at the interface are inevitable due to the phys-

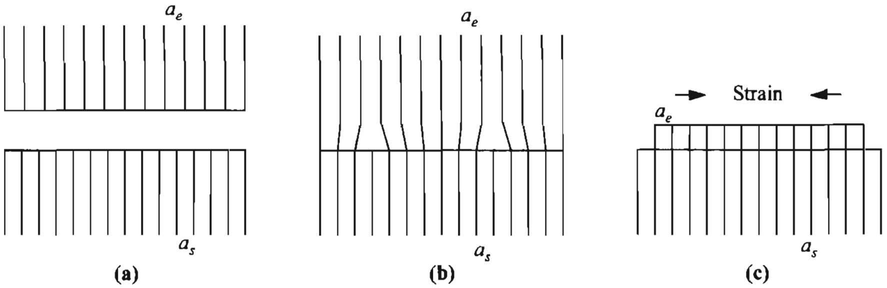


Fig. 33 Two materials with slightly mismatched lattice constants a_s and a_e. (a) In isolation. (b) Heteroepitaxy with thick, relaxed epitaxial layer having dislocations at the interface. (c) With thin, strained epitaxial layer without dislocations. Epitaxial lattice constant a_e is strained to follow that of the substrate a_s.

ical mismatch of terminating bonds at the interface. However, if the heteroepitaxial layer is thin enough, the layer can be physically strained to the degree that its lattice constant becomes the same as the substrate (Fig. 33c). When that happens, dislocations can be eliminated.

To estimate the critical thickness of this strained layer, we visualize the heteroepitaxial process from the beginning. At the start, the epitaxial layer follows the lattice of the substrate, but the strain energy builds up as the film becomes thicker. Eventually the film has built up too much strain to sustain and it transforms to a relaxed state, i.e. going from Figs. 33c to 33b. The lattice mismatch is defined as

$$\Delta \equiv \frac{|a_e - a_s|}{a_e}, \tag{140}$$

where a_e and a_s are the lattice constants of the epitaxial layer and substrate respectively. The critical thickness has been found to follow an empirical formula given by

$$t_c \approx \frac{a_e}{2\Delta} \approx \frac{a_e^2}{2|a_e - a_s|}. \tag{141}$$

A typical number for the critical thickness, from a mismatch of 2% and an a_e of 5 Å, is about 10 nm. This technique of growing strained heteroepitaxy has bough an extra degree of freedom and permits the use of a wider range of materials. It has had great impacts on expanding the applications of heterostructures, for making novel devices as well as improving their performances.

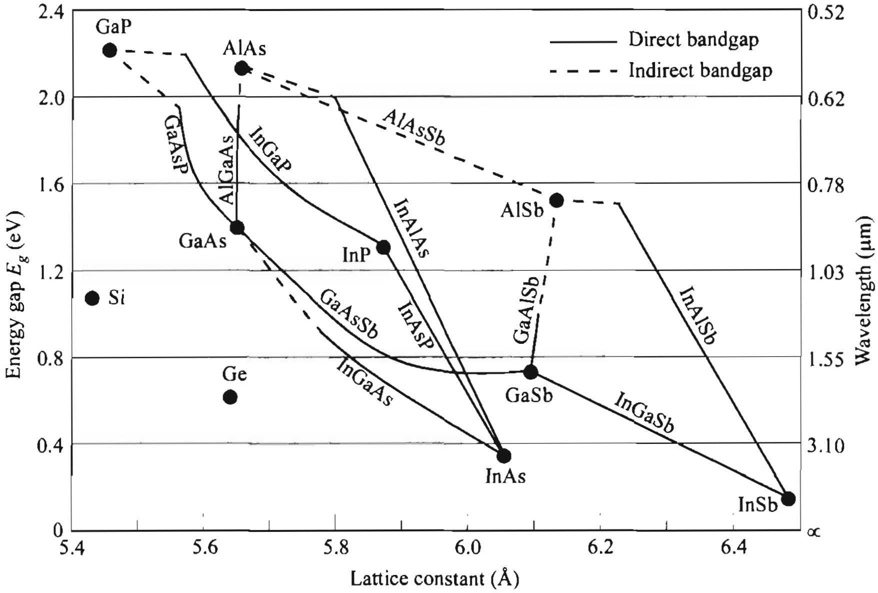


Fig. 32 Energy gap vs. lattice constant for some common elementary and binary semiconductors.

In addition to having different energy gaps, the electron affinities of these semiconductors are also different and need to be considered in device applications. This leads to different combinations of E_C and E_V alignment at the interface. According to their band alignment, heterojunctions can be classified into three groups as shown in Fig. 34: (1) Type-I or *straddling* heterojunction, (2) Type-II or *staggered* heterojunction, and (3) Type-III or *broken-gap* heterojunction. In a Type-I (straddling) heterojunction, one material has both lower E_C and higher E_V, and naturally it must have a smaller energy gap. In a Type-II (staggered) heterojunction, the locations of lower E_C and higher E_V are displaced, so electrons being collected at lower E_C and holes being collected at higher E_V are confined in different spaces. A Type-III (broken-gap) heterojunction is a special case of Type-II, but the E_C of one side is lower than the E_V of the other. The conduction band thus overlaps the valence band at the interface, hence the name *broken gap*.

Quantum Well and Superlattice. One important application of heterojunction is to use ΔE_C and ΔE_V to form barriers for carriers. A *quantum well* is formed by two heterojunctions or three layers of materials such that the middle layer has the lowest E_C for an electron well or the highest E_V for a hole well. A quantum well thus confines electrons or holes in a two-dimensional (2-D) system. When electrons are free to move in a bulk semiconductor in all directions (3-D), their energy above the conduction-band edge is continuous, given by the relationship to their momentum (Eq. 8):

$$E - E_C = \frac{\hbar^2}{2m_e^*}(k_x^2 + k_y^2 + k_z^2). \tag{142}$$

In a quantum well, carriers are confined in one direction, say in the x-coordinate such that $k_x = 0$. It will be shown that the energy within this well is no longer continuous with respect to the x-direction, but becomes quantized in subbands.

The most-important parameters for a quantum well are the well width L_x and well height ϕ_b. The energy-band diagram in Fig. 35a shows that the potential barrier is obtained from the conduction-band and valence-band offsets (ΔE_C and ΔE_V). The solution for the wavefunction of the Schrödinger equation inside the well is

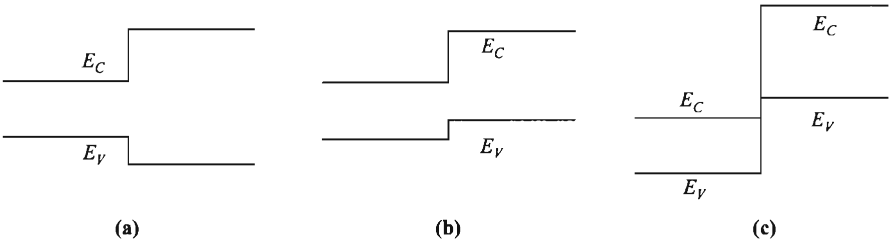


Fig. 34 Classification of heterojunctions. (a) Type-I or straddling heterojunction. (b) Type-II or staggered heterojunction. (c) Type-III or broken-gap heterojunction.

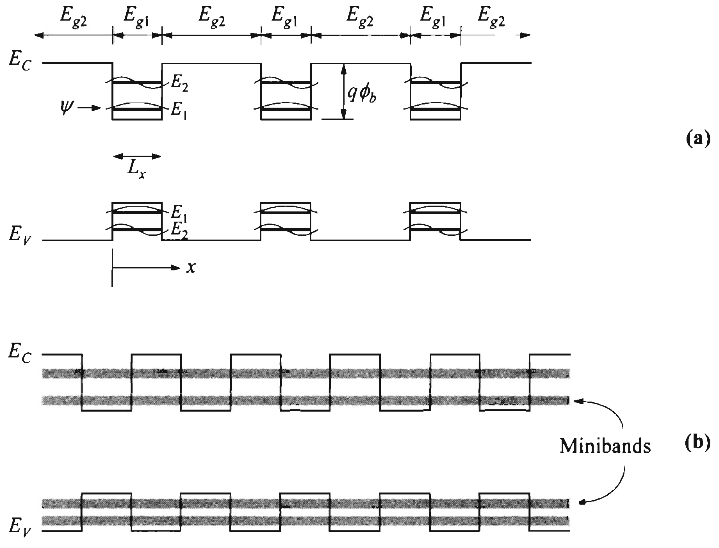


Fig. 35 Energy-band diagrams for (a) heterostructure (composition) multiple quantum wells and (b) heterostructure superlattice.

$$\psi(x) = \sin\left(\frac{i\pi x}{L_x}\right) \tag{143}$$

where i is an integer. It should be noted that at the well boundaries, ψ is truly zero only when ϕ_b is infinite. With finite ϕ_b, carriers can "leak" out (by tunneling) of the well with finite probability. This is important for the formation of a superlattice, discussed later. The pinning of nodes at the well boundaries leads to the quantization of subbands, each has a bottom energy of (with respect to the band edges)

$$E_i = \frac{\hbar^2\pi^2 i^2}{2m^* L_x^2} . \tag{144}$$

These solutions do not take into account a finite barrier height. With L_x as a variable, a quantum well can only be loosely defined. The minimum requirements should be that the quantized energy $\hbar^2\pi^2/2m^*L_x^2$ is much larger than kT, and L_x is smaller than the mean free path and the de Broglie wavelength. (Notice that the de Broglie wavelength $\lambda = h/(2m^*E)^{1/2}$ has a form similar to L_x of Eq. 144.) Also, it is interesting to note that since the continuous conduction band is now divided into subbands, carriers no longer reside on the band edges E_C or E_V but on these subbands only. In effect, the effective energy gap for interband transitions inside the quantum well becomes larger than the bulk E_g.

When quantum wells are separated from one another by thick barrier layers, there is no communication between them and this system can only be described as multiple quantum wells. However, when the barrier layers between them become thinner, to the extent that wavefunctions start to overlap, a heterostructure (composition) *superlattice* is formed. The superlattice has two major differences from a multiple-quantum-well system: (1) the energy levels are continuous in space across the barrier, and (2) the discrete bands widen into minibands (Fig. 35b). The transition from multiple quantum wells into a superlattice is analogous to the formation of a regular lattice by pulling atoms together. The isolated atoms have discrete levels, whereas a lattice transforms these discrete levels into the continuous conduction band and valance band.

Another approach to form quantum wells and superlattices is by spatial variation in doping,[84] where the potential barriers are formed by space-charge fields (Fig. 36a). The barrier shape in this case is parabolic rather than rectangular. There are two interesting features in this doping (or *n-i-p-i*) multiple-quantum-well structure. First, the conduction-band minimum and the valence-band maximum are displaced from each other, meaning that electrons and holes accumulate at different locations. This leads to minimal electron-hole recombination and very long carrier lifetime, many orders of magnitude higher than that of the regular material. This is similar to a Type-II heterojunction. Second, the effective energy gap, which is now between the first quantized levels for the electrons and holes, is reduced from the fundamental material. This tunable effective energy gap enables light emission and absorption of longer wavelengths. This structure is unique in that it has an indirect energy gap in "real space", as opposed to *k*-space. When the doping quantum wells are close together, a doping (*n-i-p-i*) superlattice is again formed (Fig. 36b).

Quantum Wire and Quantum Dot. The physical dimensions of a semiconductor have significant implications on the electronic properties, as these dimensions are

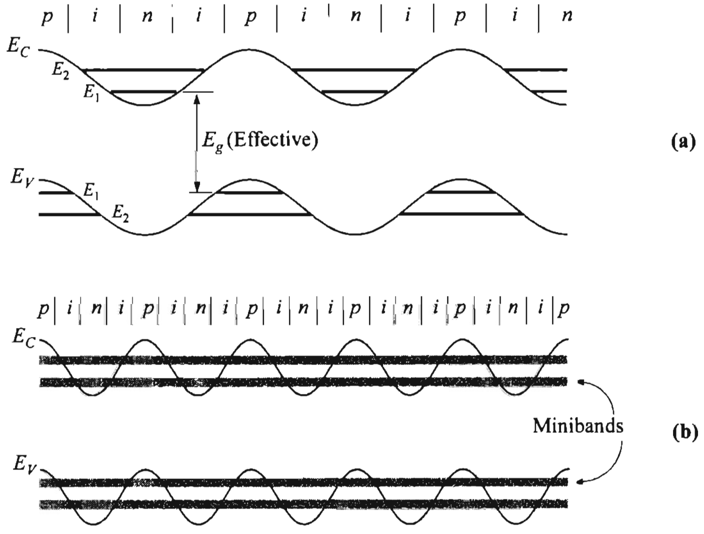


Fig. 36 Energy-band diagrams for (a) doping (*n-i-p-i*) multiple quantum wells, and (b) doping superlattice.

reduced to the order of the de Broglie wavelength. The confinement of carriers can be further extended to one- and zero-dimension, resulting in what are known as *quantum wire* and *quantum dot*. One of the major effects is on the density of states $N(E)$. Depending on the degree of confinement, $N(E)$ has very different shapes as a function of energy. The qualitative shapes of $N(E)$ for bulk semiconductor, quantum well, quantum wire, and quantum dot are shown in Fig. 37. For a 3-D system, the density of states has been given earlier (Eq. 14) and is repeated here

$$N(E) = \frac{m^* \sqrt{2m^* E}}{\pi^2 \hbar^3} . \tag{145}$$

The density of states in a 2-D system (quantum well) has a step function of

$$N(E) = \frac{m^* i}{\pi \hbar^2 L_x} . \tag{146}$$

The density of states in a 1-D system (quantum wire) has an inverse energy relationship of

$$N(E) = \frac{\sqrt{2m^*}}{\pi \hbar L_x L_y} \sum_{i,j} (E - E_{i,j})^{-1/2} , \tag{147}$$

where

$$E_{i,j} = \frac{\hbar^2 \pi^2}{2m^*} \left(\frac{i^2}{L_x^2} + \frac{j^2}{L_y^2} \right) . \tag{148}$$

The density of states in a 0-D system (quantum dot) is continuous and independent of energy,

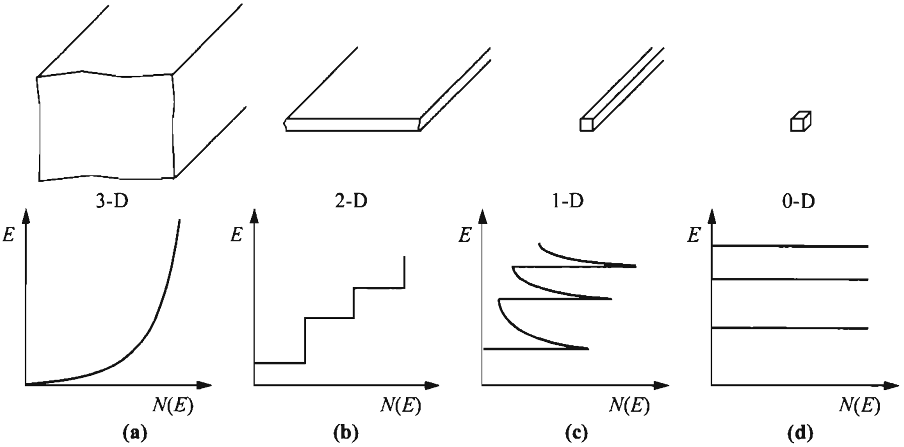


Fig. 37 Density of states $N(E)$ for (a) bulk semiconductor (3-D), (b) quantum well (2-D), (c) quantum wire (1-D), and (d) quantum dot (0-D).

$$N(E) = \frac{2}{L_x L_y L_z} \sum_{i,j,k} \delta(E - E_{i,j,k}), \tag{149}$$

where

$$E_{i,j,k} = \frac{\hbar^2 \pi^2}{2m^*}\left(\frac{i^2}{L_x^2} + \frac{j^2}{L_y^2} + \frac{k^2}{L_z^2}\right). \tag{150}$$

Since the carrier concentration and its distribution in energy is given by the density of states multiplied by the Fermi-Dirac distribution, these density-of-state functions are important for the device operation as their physical dimensions are scaled to near the de Broglie wavelength (≈ 20 nm).

1.8 BASIC EQUATIONS AND EXAMPLES

1.8.1 Basic Equations

The basic equations for semiconductor-device operation describe the static and dynamic behavior of carriers in semiconductors under external influences, such as applied field or optical excitation, that cause deviation from the thermal-equilibrium conditions.[36] The basic equations can be classified in three groups; electrostatic equations, current-density equations, and continuity equations.

Electrostatic Equations. There are two important equations relating charge to electric field (= $\mathscr{D}/\varepsilon_s$ where $\mathscr{D}$ is electric displacement). The first is from one of Maxwell equations,

$$\nabla \cdot \mathscr{D} = \rho(x, y, z), \tag{151}$$

also known as Gauss' law or Poisson equation. For a one-dimensional problem, this reduces to a more useful form of

$$\frac{d^2 \psi_i}{dx^2} = -\frac{d\mathscr{E}}{dx} = -\frac{\rho}{\varepsilon_s} = \frac{q(n - p + N_A - N_D)}{\varepsilon_s} \tag{152}$$

($\psi_i \equiv -E_i/q$). This is commonly used, for example, to determine the potential and field distribution caused by a charge density ρ within the depletion layer. The second equation deals with charge density along an interface, instead of bulk charge. The boundary conditions across an interface of charge sheet Q is given by

$$\mathscr{E}_1(0^-)\varepsilon_1 = \mathscr{E}_2(0^+)\varepsilon_2 - Q. \tag{153}$$

Current-Density Equations. The most-common current conduction consists of the drift component, caused by the electric field, and the diffusion component, caused by the carrier-concentration gradient. The current-density equations are:

$$J_n = q\mu_n n\mathscr{E} + qD_n \nabla n, \tag{154a}$$

$$J_p = q\mu_p p\mathscr{E} - qD_p \nabla p, \tag{154b}$$

$$J_{\text{cond}} = J_n + J_p, \tag{155}$$

where $\boldsymbol{J}_n$ and $\boldsymbol{J}_p$ are the electron and hole current densities, respectively. The values of the electron and hole mobilities (μ_n and μ_p) have been given in Section 1.5.1. For nondegenerate semiconductors the carrier diffusion constants (D_n and D_p) and the mobilities are given by the Einstein relation [$D_n = (kT/q)\mu_n$, etc.].

For a one-dimensional case, Eqs. 154a and 154b reduce to

$$J_n = q\mu_n n\mathscr{E} + qD_n\frac{dn}{dx} = q\mu_n\left(n\mathscr{E} + \frac{kT}{q}\frac{dn}{dx}\right) = \mu_n n\frac{dE_{Fn}}{dx}, \tag{156a}$$

$$J_p = q\mu_p p\mathscr{E} - qD_p\frac{dp}{dx} = q\mu_p\left(p\mathscr{E} - \frac{kT}{q}\frac{dp}{dx}\right) = \mu_p p\frac{dE_{Fp}}{dx} \tag{156b}$$

where E_{Fn} and E_{Fp} are quasi Fermi levels for electrons and holes, respectively. These equations are valid for low electric fields. At sufficiently high fields the term $\mu_n\mathscr{E}$ or $\mu_p\mathscr{E}$ should be replaced by the saturation velocity v_s (and the last equalities about E_{Fn} and E_{Fp} do not hold any more). These equations do not include the effect from an externally applied magnetic field where the magneto-resistive effect reduces the current.

Continuity Equations. While the above current-density equations are for steady-state conditions, the continuity equations deal with time-dependent phenomena such as low-level injection, generation and recombination. Qualitatively, the net change of carrier concentration is the difference between generation and recombination, plus the net current flowing in and out of the region of interest. The continuity equations are:

$$\frac{\partial n}{\partial t} = G_n - U_n + \frac{1}{q}\nabla\cdot\boldsymbol{J}_n, \tag{157a}$$

$$\frac{\partial p}{\partial t} = G_p - U_p - \frac{1}{q}\nabla\cdot\boldsymbol{J}_p \tag{157b}$$

where G_n and G_p are the electron and hole generation rate (cm^{-3}-s^{-1}), respectively, caused by external influences such as the optical excitation with photons or impact ionization under large electric fields. The recombination rates, $U_n = \Delta n/\tau_n$ and $U_p = \Delta p/\tau_p$, have been discussed in Section 1.5.4.

For the one-dimensional case under a low-injection condition, Eqs. 157a and 157b reduce to

$$\frac{\partial n_p}{\partial t} = G_n - \frac{n_p - n_{po}}{\tau_n} + n_p\mu_n\frac{\partial\mathscr{E}}{\partial x} + \mu_n\mathscr{E}\frac{\partial n_p}{\partial x} + D_n\frac{\partial^2 n_p}{\partial x^2} \tag{158a}$$

$$\frac{\partial p_n}{\partial t} = G_p - \frac{p_n - p_{no}}{\tau_p} - p_n\mu_p\frac{\partial\mathscr{E}}{\partial x} - \mu_p\mathscr{E}\frac{\partial p_n}{\partial x} + D_p\frac{\partial^2 p_n}{\partial x^2}. \tag{158b}$$

1.8.2 Examples

In this section, we demonstrate the use of the continuity equations for studying the time dependence and space dependence of excess carriers. Excess carriers can be created by optical excitation or injection from a nearby junction. In these examples we use optical excitation for simplicity.

Decay of Excess Carriers with Time. Consider an n-type sample, as shown in Fig. 38a, that is illuminated with light in which the electron-hole pairs are generated uniformly throughout the sample with a uniform generation rate G_p. In this example the sample thickness is much smaller than $1/\alpha$, and the space dependence is absent here. The boundary conditions are $\mathscr{E} = \partial\mathscr{E}/\partial x = 0$ and $\partial p_n/\partial x = 0$. We have from Eq. 158b:

$$\frac{dp_n}{dt} = G_p - \frac{p_n - p_{no}}{\tau_p}. \tag{159}$$

At steady state, $\partial p_n/\partial t = 0$ and

$$p_n - p_{no} = \tau_p G_p = \text{constant}. \tag{160}$$

If at an arbitrary time, say $t = 0$, the light is suddenly turned off, the differential equation is now

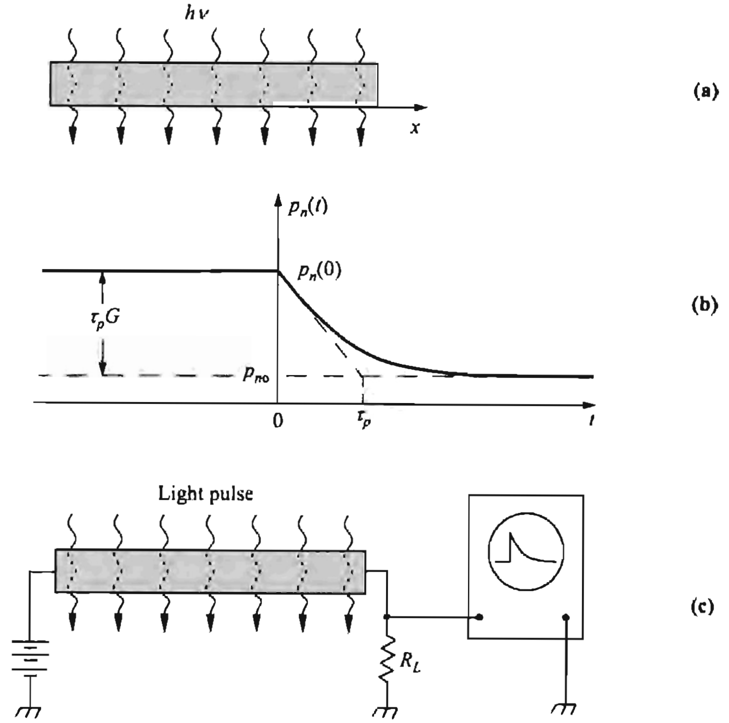


Fig. 38 Decay of photo-excited carriers. (a) n-type sample under constant illumination. (b) Decay of minority carriers (holes) with time. (c) Schematic experimental setup to measure minority carrier lifetime. (After Ref. 71.)

$$\frac{dp_n}{dt} = -\frac{p_n - p_{no}}{\tau_p}. \tag{161}$$

With the boundary conditions $p_n(t=0) = p_{no} + \tau_p G_p$, as given in Eq. 160, and $p_n(\infty) = p_{no}$, the solution is

$$p_n(t) = p_{no} + \tau_p G_p \exp\left(-\frac{t}{\tau_p}\right). \tag{162}$$

Figure 38b shows the variation of p_n with time.

The example above presents the main idea of the Stevenson-Keyes method for measuring minority-carrier lifetime.[71] Figure 38c shows a schematic setup. The excess carriers generated uniformly throughout the sample by the light pulses cause a momentary increase in the conductivity and current. During the periods when the light pulses are off, the decay of this photoconductivity can be observed on an oscilloscope which monitors the voltage drop across a resistor load R_L, and is a measure of the lifetime.

Decay of Excess Carriers with Distance. Figure 39a shows another simple example where excess carriers are injected from one side (e.g., by high-energy photons that create electron-hole pairs at the surface only). Referring to Fig. 30, note that for $h\nu = 3.5$ eV, the absorption coefficient is about 10^6 cm^{-1}, in other words, the light intensity decreases by a factor of e in a distance of 10 nm.

At steady state there is a concentration gradient near the surface. The differential equation for an n-type sample without bias is, from Eq. 158b,

$$\frac{\partial p_n}{\partial t} = 0 = -\frac{p_n - p_{no}}{\tau_p} + D_p \frac{\partial^2 p_n}{\partial x^2}. \tag{163}$$

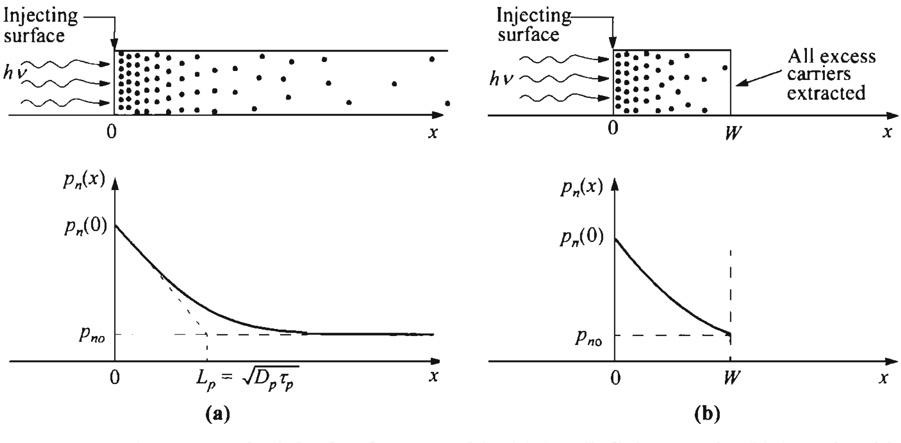


Fig. 39 Steady-state carrier injection from one side. (a) Semiinfinite sample. (b) Sample with length W.

The boundary conditions are $p_n(x = 0)$ = constant, depending on the injection level, and $p_n(\infty) = p_{no}$. The solution of $p_n(x)$ is

$$p_n(x) = p_{no} + [p_n(0) - p_{no}]\exp\left(-\frac{x}{L_p}\right) \tag{164}$$

where the diffusion length is $L_p = (D_p\tau_p)^{1/2}$ (Fig. 39a). The maximum values of L_p and L_n are of the order of 1 cm in silicon, but only of the order of 10^{-2} cm in gallium arsenide.

Of special interest is the case where the second boundary condition is changed so that all excess carriers at the back surface ($x = W$) are extracted or $p_n(W) = p_{no}$, then we obtain from Eq. 163 a new solution,

$$p_n(x) = p_{no} + [p_n(0) - p_{no}]\left\{\frac{\sinh[(W-x)/L_p]}{\sinh(W/L_p)}\right\}. \tag{165}$$

This result is shown in Fig. 39b. The current density at $x = W$ is given by Eq. 156b:

$$J_p = -qD_p\frac{dp}{dx}\bigg|_W = \frac{qD_p[p_n(0) - p_{no}]}{L_p\sinh(W/L_p)}. \tag{166}$$

It will be shown later that Eq. 166 is related to the current gain in bipolar transistors (Chapter 5).

Decay of Excess Carriers with Time and Distance. When localized light pulses generate excess carriers in a semiconductor (Fig. 40a), the transport equation after the pulse without bias is given by Eq. 158b by setting $G_p = \mathscr{E} = \partial\mathscr{E}/\partial x = 0$:

$$\frac{\partial p_n}{\partial t} = -\frac{p_n - p_{no}}{\tau_p} + D_p\frac{\partial^2 p_n}{\partial x^2}. \tag{167}$$

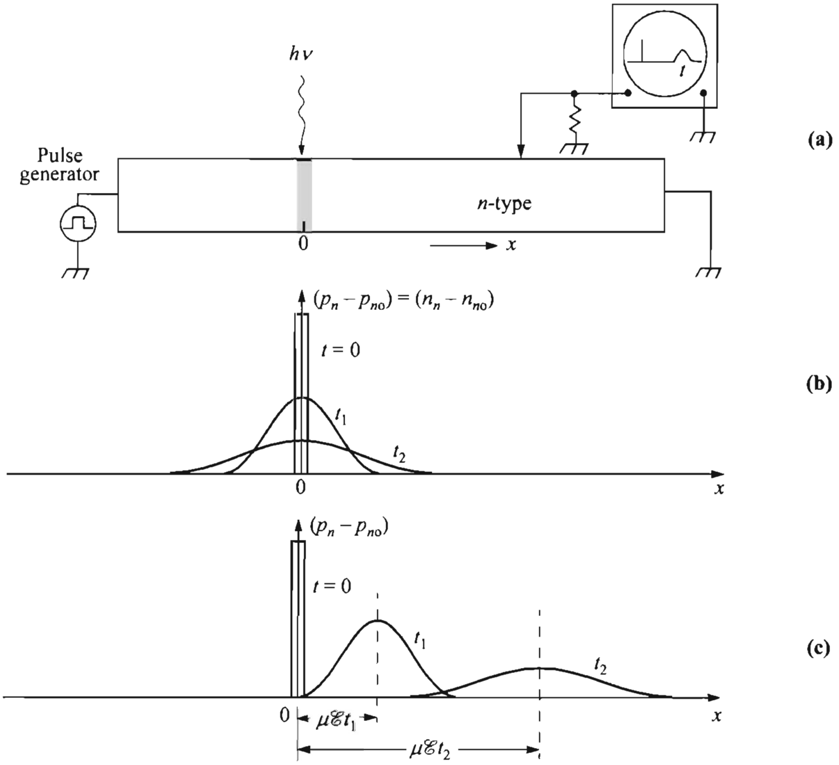
The solution is given by

$$p_n(x, t) = \frac{N'}{\sqrt{4\pi D_p t}}\exp\left(-\frac{x^2}{4D_p t} - \frac{t}{\tau_p}\right) + p_{no} \tag{168}$$

where N' is the number of electrons or holes generated initially per unit area. Figure 40b shows this solution as the carriers diffuse away from the point of injection, and they also recombine (area under curve is decreased).

If an electric field is applied along the sample, the solution is in the same form but with x replaced by $(x - \mu_p\mathscr{E}t)$ (Fig. 40c); thus the whole *package* of excess carrier moves toward the negative end of the sample with a drift velocity $\mu_p\mathscr{E}$. At the same time, the carriers diffuse outward and recombine as in the field-free case.

The example above is similar to the celebrated Haynes-Shockley experiment for the measurement of carrier drift mobility in semiconductors.[85] With known sample length, applied field, and the time delay between the applied signals (bias on and light off) and the detected signal at the sample end (both displayed on the oscilloscope), the drift mobility $\mu = x/\mathscr{E}t$ can be calculated.



hν
Pulse generator
n-type
0
x
(a)
$(p_n - p_{no}) = (n_n - n_{no})$
t = 0
t_1
t_2
0
x
(b)
$(p_n - p_{no})$
t = 0
t_1
t_2
0
x
$\mu\mathscr{E}t_1$
$\mu\mathscr{E}t_2$
(c)

Fig. 40 Transient and steady-state carrier diffusion after a localized light pulse. (a) Experimental setup. (b) Without applied field. (c) With applied field.

Surface Recombination. When surface recombination is introduced at one end of a semiconductor sample (Fig. 41), the boundary condition at $x = 0$ is governed by

$$qD_p \frac{dp_n}{dx}\bigg|_{x=0} = qS_p[p_n(0) - p_{no}] \tag{169}$$

which states that the minority carriers that reach the surface recombine there. The constant S_p with units cm/s is defined as the surface recombination velocity for holes. The boundary condition at $x = \infty$ is given by Eq. 160. The differential equation, without bias and at steady state, is

$$0 = G_p - \frac{p_n - p_{no}}{\tau_p} + D_p \frac{d^2 p_n}{dx^2}. \tag{170}$$

The solution of the equation subject to the boundary conditions above is

$$p_n(x) = p_{no} + \tau_p G_p \left[1 - \frac{\tau_p S_p \exp(-x/L_p)}{L_p + \tau_p S_p}\right] \tag{171}$$

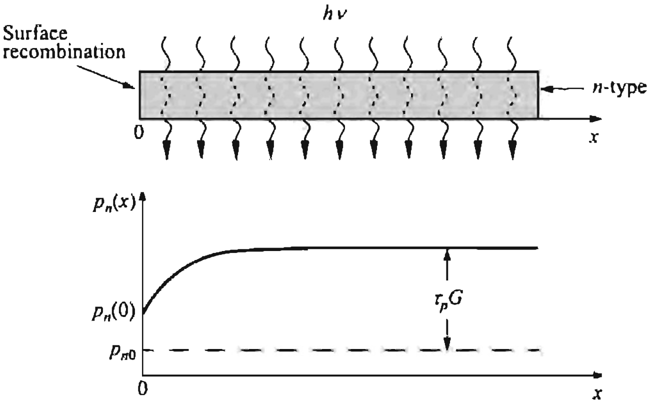


Fig. 41 Surface recombination at $x = 0$. The minority-carrier distribution near the surface is affected by the surface recombination velocity.

which is plotted in Fig. 41 for a finite S_p. When $S_p \to 0$, then $p_n(x) \to p_{no} + \tau_p G_p$, which was obtained previously (Eq. 160). When $S_p \to \infty$, then $p_n(x) \to p_{no} + \tau_p G_p[1 - \exp(-x/L_p)]$, and the minority carrier density at the surface approaches its thermal equilibrium value p_{no}. Analogous to the low-injection bulk-recombination process, in which the reciprocal of the minority-carrier lifetime $(1/\tau)$ is equal to $\sigma_p v_{th} N_t$ (Eq. 95a), the surface recombination velocity is given by

$$S_p = \sigma_p v_{th} N'_{st} \tag{172}$$

where N'_{st} is the number of surface trapping centers per unit area at the boundary region.

REFERENCES

1. W. C. Dunlap, *An Introduction to Semiconductors*, Wiley, New York, 1957.
2. O. Madelung, *Physics of III-V Compounds*, Wiley, New York, 1964.
3. J. L. Moll, *Physics of Semiconductors*, McGraw-Hill, New York, 1964.
4. T. S. Moss, Ed., *Handbook on Semiconductors*, Vols. 1–4, North-Holland, Amsterdam, 1980.
5. R. A. Smith, *Semiconductors*, 2nd Ed., Cambridge University Press, London, 1979.
6. K. W. Böer, *Survey of Semiconductor Physics*, Van Nostrand Reinhold, New York, 1990.
7. K. Seeger, *Semiconductor Physics*, 7th Ed., Springer-Verlag, Berlin, 1999.
8. S. Wang, *Fundamentals of Semiconductor Theory and Device Physics*, Prentice-Hall, Englewood Cliffs, New Jersey, 1989.
9. R. K. Willardson and A. C. Beer, Eds., *Semiconductors and Semimetals*, Vol. 2, *Physics of III-V Compounds*, Academic, New York, 1966.

10. W. B. Pearson, *Handbook of Lattice Spacings and Structure of Metals and Alloys*, Pergamon, New York, 1967.
11. H. C. Casey, Jr. and M. B. Panish, *Heterostructure Lasers*, Academic, New York, 1978.
12. See, for example, C. Kittel, *Introduction to Solid State Physics*, 7th Ed., Wiley, New York, 1996.
13. L. Brillouin, *Wave Propagation in Periodic Structures*, 2nd Ed., Dover, New York, 1963.
14. J. M. Ziman, *Principles of the Theory of Solids*, Cambridge University Press, London, 1964.
15. M. L. Cohen, "Pseudopotential Calculations for II-VI Compounds," in D. G. Thomas, Ed., *II-VI Semiconducting Compounds*, W. A. Benjamin, New York, 1967, p. 462.
16. C. Kittel, *Quantum Theory of Solids*, Wiley, New York, 1963.
17. L. C. Allen, "Interpolation Scheme for Energy Bands in Solids," *Phys. Rev.*, **98**, 993 (1955).
18. F. Herman, "The Electronic Energy Band Structure of Silicon and Germanium," *Proc. IRE*, **43**, 1703 (1955).
19. J. C. Phillips, "Energy-Band Interpolation Scheme Based on a Pseudopotential," *Phys. Rev.*, **112**, 685 (1958).
20. M. L. Cohen and J. R. Chelikowsky, *Electronic Structure and Optical Properties of Semiconductors*, 2nd Ed., Springer-Verlag, Berlin, 1988.
21. J. M. Ziman, *Electrons and Phonons*, Clarendon, Oxford, 1960.
22. C. D. Thurmond, "The Standard Thermodynamic Function of the Formation of Electrons and Holes in Ge, Si, GaAs and GaP," *J. Electrochem. Soc.*, **122**, 1133 (1975).
23. V. Alex, S. Finkbeiner, and J. Weber, "Temperature Dependence of the Indirect Energy Gap in Crystalline Silicon," *J. Appl. Phys.*, **79**, 6943 (1996).
24. W. Paul and D. M. Warschauer, Eds., *Solids under Pressure*, McGraw-Hill, New York, 1963.
25. R. S. Ohl, "Light-Sensitive Electric Device," U.S. Patent 2,402,662. Filed May 27, 1941. Granted June 25, 1946.
26. M. Riordan and L. Hoddeson, "The Origins of the *pn* Junction", *IEEE Spectrum*, **34**-6, 46 (1997).
27. J. S. Blackmore, "Carrier Concentrations and Fermi Levels in Semiconductors," *Electron. Commun.*, **29**, 131 (1952).
28. W. B. Joyce and R. W. Dixon, "Analytic Approximations for the Fermi Energy of an Ideal Fermi Gas," *Appl. Phys. Lett.*, **31**, 354 (1977).
29. O. Madelung, Ed., *Semiconductors–Basic Data*, 2nd Ed., Springer-Verlag, Berlin, 1996.
30. R. N. Hall and J. H. Racette, "Diffusion and Solubility of Copper in Extrinsic and Intrinsic Germanium, Silicon, and Gallium Arsenide," *J. Appl. Phys.*, **35**, 379 (1964).
31. A. G. Milnes, *Deep Impurities in Semiconductors*, Wiley, New York, 1973.
32. J. Hermanson and J. C. Phillips, "Pseudopotential Theory of Exciton and Impurity States," *Phys. Rev.*, **150**, 652 (1966).
33. J. Callaway and A. J. Hughes, "Localized Defects in Semiconductors," *Phys. Rev.*, **156**, 860 (1967).
34. E. M. Conwell, "Properties of Silicon and Germanium, Part II," *Proc. IRE*, **46**, 1281 (1958).

35. S. M. Sze and J. C. Irvin, "Resistivity, Mobility, and Impurity Levels in GaAs, Ge, and Si at 300 K," *Solid-State Electron.*, **11**, 599 (1968).
36. W. Shockley, *Electrons and Holes in Semiconductors*, D. Van Nostrand, Princeton, New Jersey, 1950.
37. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.
38. J. Bardeen and W. Shockley, "Deformation Potentials and Mobilities in Nonpolar Crystals," *Phys. Rev.*, **80**, 72 (1950).
39. E. Conwell and V. F. Weisskopf, "Theory of Impurity Scattering in Semiconductors," *Phys. Rev.*, **77**, 388 (1950).
40. C. Bulucea, "Recalculation of Irvin's Resistivity Curves for Diffused Layers in Silicon Using Updated Bulk Resistivity Data," *Solid-State Electron.*, **36**, 489 (1993).
41. C. Jacoboni, C. Canali, G. Ottaviani, and A. A. Quaranta, "A Review of Some Charge Transport Properties of Silicon," *Solid-State Electron.*, **20**, 77 (1977).
42. W. E. Beadle, J. C. C. Tsai, and R. D. Plummer, Eds., *Quick Reference Manual for Silicon Integrated Circuit Technology*, Wiley, New York, 1985.
43. F. M. Smits, "Measurement of Sheet Resistivities with the Four-Point Probe," *Bell Syst. Tech. J.*, **37**, 711 (1958).
44. E. H. Hall, "On a New Action of the Magnet on Electric Currents," *Am. J. Math.*, **2**, 287 (1879).
45. L. J. Van der Pauw, "A Method of Measuring Specific Resistivity and Hall Effect of Disc or Arbitrary Shape," *Philips Res. Rep.*, **13**, 1 (Feb. 1958).
46. D. M. Caughey and R. E. Thomas, "Carrier Mobilities in Silicon Empirically Related to Doping and Field," *Proc. IEEE*, **55**, 2192 (1967).
47. D. E. Aspnes, "GaAs Lower Conduction-Band Minima: Ordering and Properties," *Phys. Rev.*, **B14**, 5331 (1976).
48. P. Smith, M. Inoue, and J. Frey, "Electron Velocity in Si and GaAs at Very High Electric Fields," *Appl. Phys. Lett.*, **37**, 797 (1980).
49. J. G. Ruch and G. S. Kino, "Measurement of the Velocity-Field Characteristics of Gallium Arsenide," *Appl. Phys. Lett.*, **10**, 40 (1967).
50. K. Brennan and K. Hess, "Theory of High-Field Transport of Holes in GaAs and InP," *Phys. Rev. B*, **29**, 5581 (1984).
51. B. Kramer and A. Mircea, "Determination of Saturated Electron Velocity in GaAs," *Appl. Phys. Lett.*, **26**, 623 (1975).
52. K. K. Thornber, "Relation of Drift Velocity to Low-Field Mobility and High Field Saturation Velocity," *J. Appl. Phys.*, **51**, 2127 (1980).
53. J. G. Ruch, "Electron Dynamics in Short Channel Field-Effect Transistors," *IEEE Trans. Electron Devices*, **ED-19**, 652 (1972).
54. K. K. Thornber, "Applications of Scaling to Problems in High-Field Electronic Transport," *J. Appl. Phys.*, **52**, 279 (1981).
55. R. A. Logan and S. M. Sze, "Avalanche Multiplication in Ge and GaAs *p-n* Junctions," Proc. Int. Conf. Phys. Semicond., Kyoto, and *J. Phys. Soc. Jpn. Suppl.*, **21**, 434 (1966).
56. W. N. Grant, "Electron and Hole Ionization Rates in Epitaxial Silicon at High Electric Fields," *Solid-State Electron.*, **16**, 1189 (1973).

57. G. H. Glover, "Charge Multiplication in Au-SiC (6H) Schottky Junction," *J. Appl. Phys.*, **46**, 4842 (1975).
58. T. P. Pearsall, F. Capasso, R. E. Nahory, M. A. Pollack, and J. R. Chelikowsky, "The Band Structure Dependence of Impact Ionization by Hot Carriers in Semiconductors GaAs," *Solid-State Electron.*, **21**, 297 (1978).
59. I. Umebu, A. N. M. M. Choudhury, and P. N. Robson, "Ionization Coefficients Measured in Abrupt InP Junction," *Appl. Phys. Lett.*, **36**, 302 (1980).
60. R. A. Logan and H. G. White, "Charge Multiplication in GaP *p-n* Junctions," *J. Appl. Phys.*, **36**, 3945 (1965).
61. T. P. Pearsall, "Impact Ionization Rates for Electrons and Holes in $Ga_{0.47}In_{0.53}As$," *Appl. Phys. Lett.*, **36**, 218 (1980).
62. T. P. Pearsall, R. E. Nahory, and M. A. Pollack, "Impact Ionization Rates for Electrons and Holes in $GaAs_{1-x}Sb_x$ Alloys," *Appl. Phys. Lett.*, **28**, 403 (1976).
63. L. W. Cook, G. E. Bulman, and G. E. Stillman, "Electron and Hole Impact Ionization Coefficients in InP Determined by Photomultiplication Measurements," *Appl. Phys. Lett.*, **40**, 589 (1982).
64. I. H. Oguzman, E. Bellotti, K. F. Brennan, J. Kolnik, R. Wang, and P. P. Ruden, "Theory of Hole Initiated Impact Ionization in Bulk Zincblende and Wurtzite GaN," *J. Appl. Phys.*, **81**, 7827 (1997).
65. M. R. Brozel and G. E. Stillman, Eds., *Properties of Gallium Arsenide*, 3rd Ed., INSPEC, London, 1996.
66. C. R. Crowell and S. M. Sze, "Temperature Dependence of Avalanche Multiplication in Semiconductors," *Appl. Phys. Lett.*, **9**, 242 (1966).
67. C. T. Sah, R. N. Noyce, and W. Shockley, "Carrier Generation and Recombination in *p-n* Junction and *p-n* Junction Characteristics," *Proc. IRE*, **45**, 1228 (1957).
68. R. N. Hall, "Electron-Hole Recombination in Germanium," *Phys. Rev.*, **87**, 387 (1952).
69. W. Shockley and W. T. Read, "Statistics of the Recombination of Holes and Electrons," *Phys. Rev.*, **87**, 835 (1952).
70. W. M. Bullis, "Properties of Gold in Silicon," *Solid-State Electron.*, **9**, 143 (1966).
71. D. T. Stevenson and R. J. Keyes, "Measurement of Carrier Lifetime in Germanium and Silicon," *J. Appl. Phys.*, **26**, 190 (1955).
72. W. W. Gartner, "Spectral Distribution of the Photomagnetic Electric Effect," *Phys. Rev.*, **105**, 823 (1957).
73. S. Wei and M. Y. Chou, "Phonon Dispersions of Silicon and Germanium from First-Principles Calculations," *Phys. Rev. B*, **50**, 2221 (1994).
74. C. Patel, T. J. Parker, H. Jamshidi, and W. F. Sherman, "Phonon Frequencies in GaAs," *Phys. Stat. Sol.* (b), **122**, 461 (1984).
75. W. C. Dash and R. Newman, "Intrinsic Optical Absorption in Single-Crystal Germanium and Silicon at 77°K and 300°K," *Phys. Rev.*, **99**, 1151 (1955).
76. H. R. Philipp and E. A. Taft, "Optical Constants of Silicon in the Region 1 to 10 eV," *Phys. Rev. Lett.*, **8**, 13 (1962).
77. D. E. Hill, "Infrared Transmission and Fluorescence of Doped Gallium Arsenide," *Phys. Rev.*, **133**, A866 (1964).

78. H. C. Casey, Jr., D. D. Sell, and K. W. Wecht, "Concentration Dependence of the Absorption Coefficient for *n*- and *p*-type GaAs between 1.3 and 1.6 eV," *J. Appl. Phys.*, **46**, 250 (1975).
79. C. Y. Ho, R. W. Powell, and P. E. Liley, *Thermal Conductivity of the Elements—A Comprehensive Review*, Am. Chem. Soc. and Am. Inst. Phys., New York, 1975.
80. M. G. Holland, "Phonon Scattering in Semiconductors from Thermal Conductivity Studies," *Phys. Rev.*, **134**, A471 (1964).
81. B. H. Armstrong, "Thermal Conductivity in SiO_2", in S. T. Pantelides, Ed., *The Physics of SiO_2 and Its Interfaces*, Pergamon, New York, 1978.
82. G. A. Slack, "Thermal Conductivity of Pure and Impure Silicon, Silicon Carbide, and Diamond," *J. Appl. Phys.*, **35**, 3460 (1964).
83. E. K. Sichel and J. I. Pankove, "Thermal Conductivity of GaN, 25–360 K," *J. Phys. Chem. Solids*, **38**, 330 (1977).
84. G. H. Dohler, "Doping Superlattices—Historical Overview", in P. Bhattacharya, Ed., *III-V Quantum Wells and Superlattices*, INSPEC, London, 1996.
85. J. R. Haynes and W. Shockley, "The Mobility and Life of Injected Holes and Electrons in Germanium," *Phys. Rev.*, **81**, 835 (1951).

PROBLEMS

1. (a) Find the maximum fraction of a conventional unit-cell volume which can be filled by identical hard spheres in a diamond lattice.
(b) Find the number of atoms per square centimeter in silicon in (111) plane at 300 K.

2. Calculate the tetrahedral bond angle, i.e., the angle between any pair of the four bonds. (*Hint*: Represent the 4 bonds as vectors of equal lengths. What must be the sum of the 4 vectors equal? Take components of this vector equation along the direction of one of these vectors.)

3. For a face centered cubic, the volume of a conventional unit cell is a^3. Find the volume of a fcc primitive unit cell with three basis vectors: $(0,0,0 \rightarrow a/2,0,a/2)$, $(0,0,0 \rightarrow a/2,a/2,0)$, and $(0,0,0 \rightarrow 0,a/2,a/2)$.

4. (a) Derive an expression for the bond length d in the diamond lattice in terms of the lattice constant a.
(b) In a silicon crystal, if a plane has intercepts at 10.86 Å, 16.29 Å, and 21.72 Å along the three Cartesian coordinates, find the Miller indices of the plane.

5. Show (a) that each vector of the reciprocal lattice is normal to a set of planes in the direct lattice, and (b) the volume of a unit cell of the reciprocal lattice is inversely proportional to the volume of a unit cell of the direct lattice.

6. Show that the reciprocal lattice of a body-centered cubic (bcc) lattice with a lattice constant a is a face-centered cubic (fcc) lattice with the side of the cubic cell to be $4\pi/a$. [*Hint*: Use a symmetric set of vectors for bcc:

$$\boldsymbol{a} = \frac{a}{2}(\boldsymbol{y}+\boldsymbol{z}-\boldsymbol{x}), \quad \boldsymbol{b} = \frac{a}{2}(\boldsymbol{z}+\boldsymbol{x}-\boldsymbol{y}), \quad \boldsymbol{c} = \frac{a}{2}(\boldsymbol{x}+\boldsymbol{y}-\boldsymbol{z})$$

where a is the lattice constant of a conventional primitive cell, and $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ are unity vectors of a Cartesian coordinate. For fcc;

$$\boldsymbol{a} = \frac{a}{2}(\boldsymbol{y}+\boldsymbol{z}), \quad \boldsymbol{b} = \frac{a}{2}(\boldsymbol{z}+\boldsymbol{x}), \quad \boldsymbol{c} = \frac{a}{2}(\boldsymbol{x}+\boldsymbol{y}) .]$$

7. Near the conduction band minima the energy can be expressed as

$$E = \frac{\hbar^2}{2}\left(\frac{k_x^2}{m_x^*} + \frac{k_y^2}{m_y^*} + \frac{k_z^2}{m_z^*}\right).$$

In Si there are six cigar-shaped minima along [100]. If the ratio of the axes of constant energy ellipsoid is 5:1, find the ratio of longitudinal effective mass m_l^* to the transverse effective mass m_t^*.

8. In the conduction band of a semiconductor, it has a lower valley at the center of the Brillouin zone, and six upper valleys at the zone boundary along [100]. If the effective mass for the lower valley is $0.1m_0$ and that for the upper valleys is $1.0m_0$, find the ratio of the effective density of states in the upper valleys to that in the lower valley.

9. Derive the density of states in the conduction band as given by Eq. 14.
(*Hint*: The wavelength λ of a standing wave is related to the length of the semiconductor L by $L/\lambda = n_x$ where n_x is an integer. The wavelength can be expressed by de Broglie hypothesis $\lambda = h/p_x$. Consider a three-dimensional cube of side L.)

10. Calculate the average kinetic energy of electrons in the conduction band of an n-type nondegenerate semiconductor. The density of states is given by Eq. 14.

11. Show that

$$N_D^+ = N_D\left[1 + 2\exp\left(\frac{E_F - E_D}{kT}\right)\right]^{-1}.$$

[*Hint*: The probability of occupancy is

$$F(E) = \left[1 + \frac{h}{g}\exp\left(\frac{E - E_F}{kT}\right)\right]^{-1}$$

where h is the number of electrons that can physically occupy the level E, and g is the number of electrons that can be accepted by the level, also called the ground-state degeneracy of the donor impurity level ($g = 2$).]

12. If a silicon sample is doped with 10^{16} phosphorous impurities/cm^3, find the ionized donor density at 77 K. Assume that the ionization energy for phosphorous donor impurities and the electron effective mass are independent of temperature. (*Hint*: First select a N_D^+ value to calculate the Fermi level, then find the corresponding N_D^+. If they don't agree, select another N_D^+ value and repeat the process until a consistent N_D^+ is obtained.)

13. Using graphic method to determine the Fermi level for a boron-doped silicon sample with an impurity concentration of 10^{15} cm^{-3} at 300 K (note $n_i = 9.65\times10^9$ cm^{-3}).

14. The Fermi-Dirac distribution function is given by $F(E) = \dfrac{1}{1 + \exp[(E - E_F)/kT]}$. The differentiation of $F(E)$ with respect to energy is $F'(E)$. Find the width of $F'(E)$, i.e., $2\left[E(\text{at } F'_{\max}) - E\left(\text{at } \frac{1}{2}F'_{\max}\right)\right]$ where $|F'_{\max}|$ is the maximum value of $F'(E)$.

15. Find the position of the Fermi level with respect to the bottom of the conduction band ($E_C - E_F$) for a silicon sample at 300 K, which is doped with 2×10^{10} cm^{-3} fully ionized donors.

16. Gold in Si has two energy levels in the bandgap: $E_C - E_A = 0.54$ eV, $E_D - E_V = 0.29$ eV. Assume the third level $E_D - E_V = 0.35$ eV is inactive. (a) What will be the state of charge of the gold levels in Si doped with high concentration of boron atoms? Why? (b) What is the effect of gold on electron and hole concentrations?

17. From Fig. 13, evaluate and determine what kind of impurity atoms is used to dope the Si sample?

18. For an *n*-type silicon sample doped with 2.86×10^{16} cm^{-3} phosphorous atoms, find the ratio of the neutral to ionized donors at 300 K. $(E_C - E_D) = 0.045$ eV.

19. (a) Assume the mobility ratio $\mu_n/\mu_p \equiv b$ in Si is a constant independent of impurity concentration. Find the maximum resistivity ρ_m in terms of the intrinsic resistivity ρ_i at 300 K. If $b = 3$ and the hole mobility of intrinsic Si is 450 cm^2/V-s, calculate ρ_i and ρ_m.
(b) Find the electron and hole concentration, mobility, and resistivity of a GaAs sample at 300 K with 5×10^{15} zinc atoms/cm^3, 10^{17} sulfur atoms/cm^3, and 10^{17} carbon atoms/cm^3.

20. The Gamma Function is defined as $\Gamma(n) \equiv \int_0^\infty x^{n-1}\exp(-x)dx$.
(a) Find $\Gamma(1/2)$, and (b) show that $\Gamma(n) = (n-1)\Gamma(n-1)$.

21. Consider a compensated *n*-type silicon at $T = 300$ K, with a conductivity of $\sigma = 16$ S/cm and an acceptor doping concentration of 10^{17} cm^{-3}. Determine the donor concentration and the electron mobility. (A compensated semiconductor is one that contains both donor and acceptor impurity atoms in the same region.)

22. Find the resistivity at 300 K for a silicon sample doped with 1.0×10^{14} cm^{-3} of phosphorous atoms, 8.5×10^{12} cm^{-3} of arsenic atoms, and 1.2×10^{13} cm^{-3} of boron atoms. Assume that the impurities are completely ionized and the mobilities are $\mu_n = 1500$ cm^2/V-s, $\mu_p = 500$ cm^2/V-s, independent of impurity concentrations.

23. A semiconductor has a resistivity of 1.0 Ω-cm, and a Hall coefficient of -1250 cm^2/Coul. Calculate the carrier density and mobility, assuming that only one type of carrier is present and the mean free time is proportional to the carrier energy, i.e., $\tau \propto E$.

24. Derive the recombination rate for indirect recombination as given by Eq. 92.
(*Hint*: Refer to Fig. 25b, the capture rate of an electron by a recombination center is proportional to $R_e \propto nN_t(1 - F)$ where n is the density of electrons in the conduction band, N_t is the density of recombination centers, F is the Fermi distribution, and $N_t(1 - F)$ is the density of unoccupied recombination centers available for electron capture.)

25. The recombination rate is given by Eq. 92. Under low injection condition, U can be expressed as $(p_n - p_{no})/\tau_r$, where τ_r is the recombination lifetime. If $\sigma_n = \sigma_p = \sigma_o$, $n_{no} = 10^{15}$ cm^{-3}, and $\tau_{ro} \equiv (\upsilon_{th}\sigma_o N_t)^{-1}$, find the values of $(E_t - E_i)$ at which the recombination lifetime τ_r becomes $2\tau_{ro}$.

26. For single-level recombination with identical electron and hole capture cross sections, find the number of trap centers per unit volume per generation rate under the condition of complete depletion of carriers. Assume that the trap centers are located at mid bandgap, $\sigma = 2\times10^{-16}$ cm^2, and $\upsilon_{th} = 10^7$ cm/s.

27. In a region of semiconductor which is completely depleted of carriers (i.e., $n \ll n_i$, $p \ll n_i$), electron-hole pairs are generated by alternate emission of electrons and of holes by the centers. Derive the average time that takes place between such emission process (assume

$\sigma_n = \sigma_p = \sigma$); also find the average time for $\sigma = 2\times10^{-16}$ cm^2, υ_{th} = 10^7 cm/s, and $E_t = E_i$. (T = 300 K).

28. For a single-level recombination process, find the average time that takes place between each recombination process in a region of a silicon sample where $n = p = 10^{13}$ cm^{-3}, $\sigma_n = \sigma_p = 2\times10^{-16}$ cm^2, $\upsilon_{th} = 10^7$ cm/s, $N_t = 10^{16}$ cm^{-3}, and $(E_t - E_i) = 5kT$.

29. (a). Derive Eq. 123.
(*Hint*: Assume a linear chain of atoms and the atoms interact only with nearest neighbors. The even-numbered atoms have mass m_1 and the odd-numbered atoms have mass m_2.)
(b) For a silicon crystal with $m_1 = m_2$ and $\sqrt{\alpha_f/m_1} = 7.63\times10^{12}$ Hz, find the optical phonon energy at the boundary of the Brillouin zone. The force constant is α_f.

30. Assume $Ga_{0.5}In_{0.5}As$ is lattice matched with InP substrate at 500°C. When the sample is cooled to 27°C, find the lattice mismatch between the layers.

31. Find the ratio of the conduction-band discontinuity of the heterojunction $Al_{0.4}Ga_{0.6}As$/GaAs to the $Al_{0.4}Ga_{0.6}As$ bandgap.

32. In a Haynes-Shockley experiment, the maximum amplitudes of the minority carriers at t_1 = 25 μs and t_2 = 100 μs differ by a factor of 10. Find the minority carrier lifetime.

33. From the expression which describes the drift and diffusion of carriers in the Haynes-Shockley experiment, find the half-width of the pulse of carriers at t = 1 s. Assume the diffusion coefficient is 10 cm^2/s.

34. Excess carriers are injected on one surface (x = 0) of a thin slide of n-type (3×10^{17} cm^{-3}) silicon with length W = 0.05 mm and extracted at the opposite surface where $p_n(W) = p_{no}$. If the carrier lifetime is 50 μs, find the portion of injected current which reaches the opposite surface by diffusion.

35. A GaAs n-type sample with $N_D = 5\times10^{15}$ cm^{-3} is illuminated. The uniformly absorbed light creates 10^{17} electron-hole pairs/cm^3-s. The lifetime τ_p is 10^{-7} s, $L_p = 1.93\times10^{-3}$ cm, the surface recombination velocity S_p is 10^5 cm/s. Find the number of holes recombining at the surface per unit surface area in unit time.

36. An n-type semiconductor has excess carrier holes 10^{14} cm^{-3}, a minority carrier lifetime 10^{-6} s in the bulk material, and a minority carrier lifetime 10^{-7} s at the surface. Assume zero applied electric field and let D_p = 10 cm^2/s. Determine the steady-state excess carrier concentration as a function of distance from the surface (x = 0) of the semiconductor.

Copyright © 2007 John Wiley & Sons, Inc.

PART II

DEVICE BUILDING BLOCKS

- **Chapter 2** ***p-n*** **Junctions**
- **Chapter 3 Metal-Semiconductor Contacts**
- **Chapter 4 Metal-Insulator-Semiconductor Capacitors**

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

2

p-n Junctions

2.1 INTRODUCTION
2.2 DEPLETION REGION
2.3 CURRENT-VOLTAGE CHARACTERISTICS
2.4 JUNCTION BREAKDOWN
2.5 TRANSIENT BEHAVIOR AND NOISE
2.6 TERMINAL FUNCTIONS
2.7 HETEROJUNCTIONS

2.1 INTRODUCTION

p-n junctions are of great importance both in modern electronic applications and in understanding other semiconductor devices. The *p-n* junction theory serves as the foundation of the physics of semiconductor devices. The basic theory of current-voltage characteristics of *p-n* junctions was established by Shockley.[1,2] This theory was then extended by Sah, Noyce, and Shockley[3], and by Moll.[4]

The basic equations presented in Chapter 1 are used to develop the ideal static and dynamic characteristics of *p-n* junctions. Departures from the ideal characteristics due to generation and recombination in the depletion layer, to high injection, and to series resistance effects are then discussed. Junction breakdown, especially that due to avalanche multiplication, is considered in detail, after which transient behavior and noise performance in *p-n* junctions are presented.

A *p-n* junction is a two-terminal device. Depending on the doping profile, device geometry, and biasing condition, a *p-n* junction can perform various terminal functions which are considered briefly in Section 2.6. The chapter closes with a discussion of an important group of devices—the heterojunctions, which are junctions formed between dissimilar semiconductors (e.g., *n*-type GaAs on *p*-type AlGaAs).

2.2 DEPLETION REGION

2.2.1 Abrupt Junction

Built-in Potential and Depletion-Layer Width. When the impurity concentration in a semiconductor changes abruptly from acceptor impurities N_A to donor impurities N_D, as shown in Fig. 1a, one obtains an abrupt junction. In particular, if $N_A \gg N_D$ (or vice versa), one obtains a one-sided abrupt p^+-n (or n^+-p) junction.

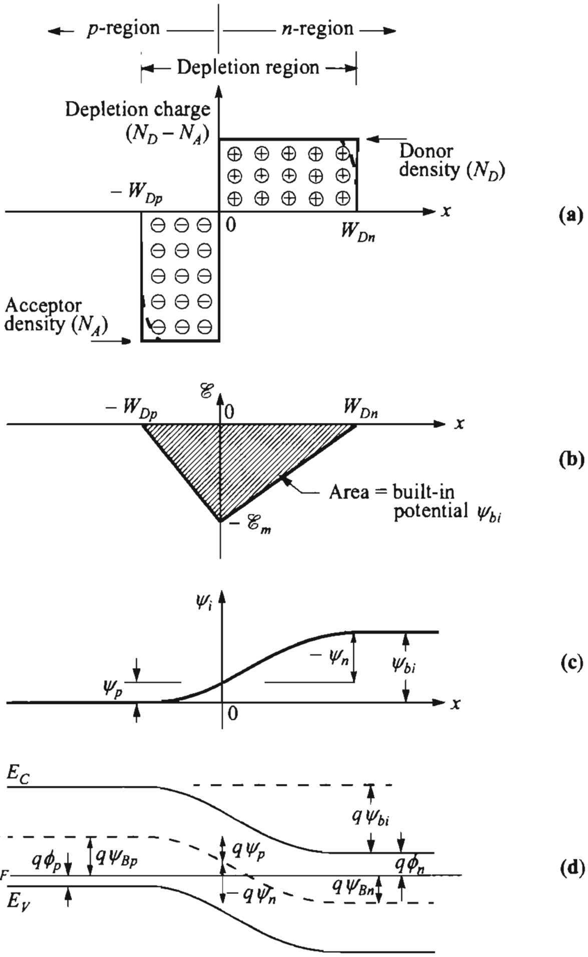

p-region
n-region
Depletion region
Depletion charge ($N_D - N_A$)
Donor density (N_D)
$-W_{Dp}$
0
W_{Dn}
x
Acceptor density (N_A)
(a)
$\mathscr{E}$
$-W_{Dp}$
0
W_{Dn}
x
Area = built-in potential ψ_{bi}
$-\mathscr{E}_m$
(b)
ψ_i
$-\psi_n$
ψ_{bi}
ψ_p
0
x
(c)
E_C
E_i
E_F
E_V
$q\phi_p$
$q\psi_{Bp}$
$q\psi_p$
$-q\psi_n$
$q\psi_{bi}$
$q\phi_n$
$q\psi_{Bn}$
(d)

Fig. 1 Abrupt *p-n* junction in thermal equilibrium. (a) Space-charge distribution. Dashed lines indicate corrections to depletion approximation. (b) Electric-field distribution. (c) Potential distribution where ψ_{bi} is the built-in potential. (c) Energy-band diagram.

We first consider the thermal equilibrium condition, that is, one without applied voltage and current flow. From the current equation of drift and diffusion (Eq. 156a in Chapter 1),

$$J_n = 0 = q\mu_n\left(n\mathscr{E} + \frac{kT}{q}\frac{dn}{dx}\right) = \mu_n n\frac{dE_F}{dx} \tag{1}$$

or

$$\frac{dE_F}{dx} = 0. \tag{2}$$

Similarly,

$$J_p = 0 = \mu_p p\frac{dE_F}{dx}. \tag{3}$$

Thus the condition of zero net electron and hole currents requires that the Fermi level must be constant throughout the sample. The built-in potential ψ_{bi}, or diffusion potential, as shown in Fig. 1b, c, and d, is equal to

$$q\psi_{bi} = E_g - (q\phi_n + q\phi_p) = q\psi_{Bn} + q\psi_{Bp}. \tag{4}$$

For nondegenerate semiconductors,

$$\psi_{bi} = \frac{kT}{q}\ln\left(\frac{n_{no}}{n_i}\right) + \frac{kT}{q}\ln\left(\frac{p_{po}}{n_i}\right)$$
$$\approx \frac{kT}{q}\ln\left(\frac{N_D N_A}{n_i^2}\right). \tag{5}$$

Since at equilibrium $n_{no}p_{no} = n_{po}p_{po} = n_i^2$,

$$\psi_{bi} = \frac{kT}{q}\ln\left(\frac{p_{po}}{p_{no}}\right) = \frac{kT}{q}\ln\left(\frac{n_{no}}{n_{po}}\right). \tag{6}$$

This gives the relationship between carrier densities on either side of the junction.

If one or both sides of the junction are degenerate, care has to be taken in calculating the Fermi-levels and built-in potential. Equation 4 has to be used since Boltzmann statistics cannot be used to simplify the Fermi-Dirac integral. Furthermore, incomplete ionization has to be considered, i.e., $n_{no} \neq N_D$ and/or $p_{po} \neq N_A$ (Eqs. 34 and 35 of Chapter 1).

Next, we proceed to calculate the field and potential distribution inside the depletion region. To simplify the analysis, the depletion approximation is used which assumes that the depleted charge has a box profile. Since in thermal equilibrium the electric field in the neutral regions (far from the junction at either side) of the semiconductor must be zero, the total negative charge per unit area in the p-side must be precisely equal to the total positive charge per unit area in the n-side:

$$N_A W_{Dp} = N_D W_{Dn}. \tag{7}$$

From the Poisson equation we obtain

$$-\frac{d^2\psi_i}{dx^2} = \frac{d\mathscr{E}}{dx} = \frac{\rho(x)}{\varepsilon_s} = \frac{q}{\varepsilon_s}[N_D^+(x) - n(x) - N_A^-(x) + p(x)] . \tag{8}$$

Inside the depletion region, $n(x) \approx p(x) \approx 0$, and assuming complete ionization,

$$\frac{d^2\psi_i}{dx^2} \approx \frac{qN_A}{\varepsilon_s} \qquad \text{for } -W_{Dp} \leq x \leq 0 , \tag{9a}$$

$$-\frac{d^2\psi_i}{dx^2} \approx \frac{qN_D}{\varepsilon_s} \qquad \text{for } 0 \leq x \leq W_{Dn} . \tag{9b}$$

The electric field is then obtained by integrating the above equations, as shown in Fig. 1b:

$$\mathscr{E}(x) = -\frac{qN_A(x + W_{Dp})}{\varepsilon_s} \qquad \text{for } -W_{Dp} \leq x \leq 0, \tag{10}$$

$$\begin{aligned} \mathscr{E}(x) &= -\mathscr{E}_m + \frac{qN_D x}{\varepsilon_s} \\ &= -\frac{qN_D}{\varepsilon_s}(W_{Dn} - x) \qquad \text{for } 0 \leq x \leq W_{Dn} \end{aligned} \tag{11}$$

where $\mathscr{E}_m$ is the maximum field that exists at $x = 0$ and is given by

$$|\mathscr{E}_m| = \frac{qN_D W_{Dn}}{\varepsilon_s} = \frac{qN_A W_{Dp}}{\varepsilon_s} . \tag{12}$$

Integrating Eqs. 10 and 11 once again gives the potential distribution $\psi_i(x)$ (Fig. 1c)

$$\psi_i(x) = \frac{qN_A}{2\varepsilon_s}(x + W_{Dp})^2 \qquad \text{for } -W_{Dp} \leq x \leq 0, \tag{13}$$

$$\psi_i(x) = \psi_i(0) + \frac{qN_D}{\varepsilon_s}\left(W_{Dn} - \frac{x}{2}\right)x \qquad \text{for } 0 \leq x \leq W_{Dn} . \tag{14}$$

With these, the potentials across different regions can be found as:

$$\psi_p = \frac{qN_A W_{Dp}^2}{2\varepsilon_s} , \tag{15a}$$

$$|\psi_n| = \frac{qN_D W_{Dn}^2}{2\varepsilon_s} , \tag{15b}$$

(ψ_n is relative to the n-type bulk and is thus negative. See definition in Appendix A)

$$\psi_{bi} = \psi_p + |\psi_n| = \psi_i(W_{Dn}) = \frac{|\mathscr{E}_m|}{2}(W_{Dp} + W_{Dn}) \tag{16}$$

where $\mathscr{E}_m$ can also be expressed as:

$$|\mathscr{E}_m| = \sqrt{\frac{2qN_A\psi_p}{\varepsilon_s}} = \sqrt{\frac{2qN_D|\psi_n|}{\varepsilon_s}} . \tag{17}$$

From Eqs. 16 and 7, the depletion widths are calculated to be:

$$W_{Dp} = \sqrt{\frac{2\varepsilon_s \psi_{bi}}{q} \frac{N_D}{N_A(N_A + N_D)}}\ , \tag{18a}$$

$$W_{Dn} = \sqrt{\frac{2\varepsilon_s \psi_{bi}}{q} \frac{N_A}{N_D(N_A + N_D)}}, \tag{18b}$$

$$W_{Dp} + W_{Dn} = \sqrt{\frac{2\varepsilon_s}{q}\left(\frac{N_A + N_D}{N_A N_D}\right)\psi_{bi}}. \tag{19}$$

The following relationships can be further deduced:

$$\frac{|\psi_n|}{\psi_{bi}} = \frac{W_{Dn}}{W_{Dp} + W_{Dn}} = \frac{N_A}{N_A + N_D}, \tag{20a}$$

$$\frac{\psi_p}{\psi_{bi}} = \frac{W_{Dp}}{W_{Dp} + W_{Dn}} = \frac{N_D}{N_A + N_D}. \tag{20b}$$

For a one-sided abrupt junction (p^+-n or n^+-p), Eq. 4 is used to calculate the built-in potential. In this case, the majority of the potential variation and depletion region will be inside the lightly doped side. Equation 19 reduces to

$$W_D = \sqrt{\frac{2\varepsilon_s \psi_{bi}}{qN}} \tag{21}$$

where N is N_D or N_A depending on whether $N_A \gg N_D$ or vice versa, and

$$\psi_i(x) = |\mathscr{E}_m|\left(x - \frac{x^2}{2W_D}\right). \tag{22}$$

This discussion uses box profiles for the depletion charges, i.e., depletion approximation. A more accurate result for the depletion-layer properties can be obtained by considering the majority-carrier contribution in addition to the impurity concentration in the Poisson equation, that is, $\rho \approx -q[N_A - p(x)]$ on the p-side and $\rho \approx q[N_D - n(x)]$ on the n-side. The depletion width is essentially the same as given by Eq. 19, except that ψ_{bi} is replaced by $(\psi_{bi} - 2kT/q)$.* The correction factor $2kT/q$ comes about because of the two majority-carrier distribution tails[5,6] (electrons in n-side and holes in p-side, as shown by the dashed lines in Fig. 1a) near the edges of the depletion region. Each contributes a correction factor kT/q. The depletion-layer width at thermal equilibrium for a one-sided abrupt junction becomes

$$W_D = \sqrt{\frac{2\varepsilon_s}{qN}\left(\psi_{bi} - \frac{2kT}{q}\right)}. \tag{23}$$

Furthermore, when a voltage V is applied to the junction, the total electrostatic potential variation across the junction is given by $(\psi_{bi} - V)$ where V is positive for forward bias (positive voltage on p-region with respect to n-region) and negative for reverse bias. Substituting $(\psi_{bi} - V)$ for ψ_{bi} in Eq. 23 yields the depletion-layer width as a function of the applied voltage. The results for one-sided abrupt junctions in

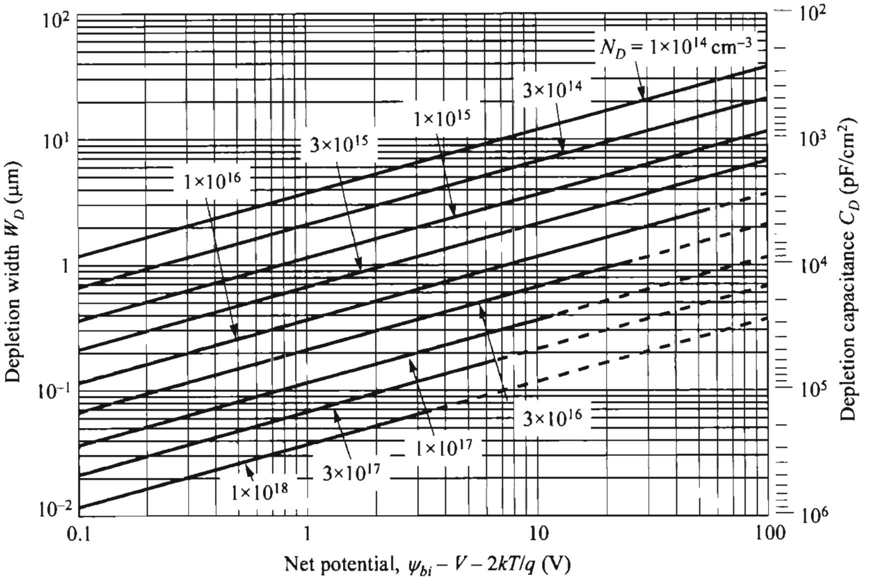


Fig. 2 Depletion-layer width and depletion-layer capacitance per unit area as a function of net potential ($\psi_{bi} - V - 2kT/q$) for one-sided abrupt junctions in Si. Doping N is from the lightly doped side. Dashed lines represent breakdown conditions.

silicon are shown in Fig. 2. The net potential at zero bias is near 0.8 V for Si and 1.3 V for GaAs. This net potential will be decreased under forward bias and increased under reverse bias. These results can also be used for GaAs since both Si and GaAs have approximately the same static dielectric constants. To obtain the depletion-layer width for other semiconductors such as Ge, one must multiply the results of Si by the factor $\sqrt{\varepsilon_s(\mathrm{Ge})/\varepsilon_s(\mathrm{Si})}$ (= 1.16). The simple model above can give adequate predictions for most abrupt *p-n* junctions.

---

* In the *p*-type region, the Poisson equation including the hole concentration is

$$\frac{d^2\psi_i}{dx^2} = \frac{q}{\varepsilon_s}[N_A - p(x)] = \frac{qN_A}{\varepsilon_s}[1 - \exp(-\beta_{th}\psi_i)].$$

Integrating both sides by $d\psi_i$, and using $d\psi_i/dx = -\mathscr{E}$,

$$\int_0^{\psi_p} -\frac{d\mathscr{E}}{dx}\,d\psi_i = \frac{qN_A}{\varepsilon_s}\int_0^{\psi_p}[1 - \exp(-\beta_{th}\psi_i)]d\psi_i,$$

$$\frac{\mathscr{E}_m^2}{2} = \frac{qN_A}{\beta_{th}\varepsilon_s}[\beta_{th}\psi_p + \exp(-\beta_{th}\psi_p) - 1] \approx \frac{qN_A}{\varepsilon_s}\left(\psi_p - \frac{kT}{q}\right).$$

Comparing this to Eq. 17, the potential is decreased by kT/q per side of the junction.

Depletion-Layer Capacitance. The depletion-layer capacitance per unit area is defined as $C_D = dQ_D/dV = \varepsilon_s/W_D$, where dQ_D is the incremental depletion charge on each side of the junction (total charge is zero) upon an incremental change of the applied voltage dV. For one-sided abrupt junctions, the capacitance per unit area is given by

$$C_D = \frac{\varepsilon_s}{W_D} = \sqrt{\frac{q\varepsilon_s N}{2}}\left(\psi_{bi} - V - \frac{2kT}{q}\right)^{-1/2} \tag{24}$$

where V is positive/negative for forward/reverse bias. The results of the depletion-layer capacitance are also shown in Fig. 2. Rearrange the above equation leads to:

$$\frac{1}{C_D^2} = \frac{2}{q\varepsilon_s N}\left(\psi_{bi} - V - \frac{2kT}{q}\right), \tag{25}$$

$$\frac{d(1/C_D^2)}{dV} = -\frac{2}{q\varepsilon_s N}. \tag{26}$$

It is clear from Eqs. 25 and 26 that by plotting $1/C^2$ versus V, a straight line should result from a one-sided abrupt junction (Fig. 3). The slope gives the impurity concentration of the substrate (N), and the extrapolation to $1/C^2 = 0$ gives ($\psi_{bi} - 2kT/q$). Note that, for the forward bias, a diffusion capacitance exists in addition to the depletion capacitance mentioned previously. The diffusion capacitance will be discussed in Section 2.3.4.

Note that the semiconductor potential and the capacitance-voltage data are insensitive to changes in the doping profiles that occur in a distance less than a Debye length.[7] The Debye length L_D is a characteristic length for semiconductors and is defined as

$$L_D \equiv \sqrt{\frac{\varepsilon_s kT}{q^2 N}} = \sqrt{\frac{\varepsilon_s}{qN\beta_{th}}}. \tag{27}$$

This Debye length gives an idea of the limit of the potential change in response to an abrupt change in the doping profile. Consider a case where the doping has a small

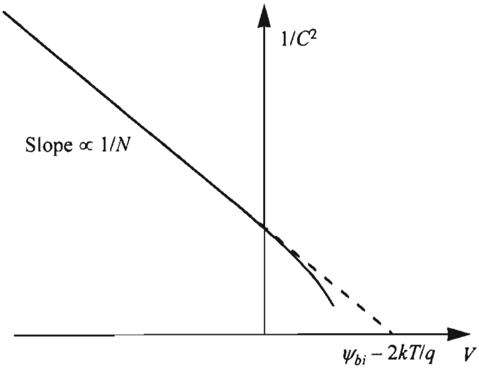


Fig. 3 A $1/C^2$-V plot can yield the built-in potential and doping density N.

increase of ΔN_D in the background of N_D, the change of potential $\Delta\psi_i(x)$ near the step is given by

$$n = N_D \exp\left(\frac{\Delta\psi_i q}{kT}\right), \tag{28}$$

$$\begin{aligned}\frac{d^2\Delta\psi_i}{dx^2} &= -\frac{q}{\varepsilon_s}(N_D+\Delta N_D-n) = -\frac{qN_D}{\varepsilon_s}\left[1+\frac{\Delta N_D}{N_D}-\exp\left(\frac{\Delta\psi_i q}{kT}\right)\right]\\ &\approx -\frac{qN_D}{\varepsilon_s}\left[1+\frac{\Delta N_D}{N_D}-\left(1+\frac{\Delta\psi_i q}{kT}\right)\right]\approx\frac{q^2N_D}{\varepsilon_s kT}\Delta\psi_i\end{aligned} \tag{29}$$

whose solution has a decay length given by Eq. 27. This implies that if the doping profile changes abruptly in a scale less than the Debye length, this variation has no effect and cannot be resolved, and that if the depletion width is smaller than the Debye length, the analysis using the Poisson equation is no longer valid. At thermal equilibrium the depletion-layer widths of abrupt junctions are about $8L_D$ for Si, and $10L_D$ for GaAs. The Debye length as a function of doping density is shown in Fig. 4 for silicon at room temperature. For a doping density of 10^{16} cm^{-3}, the Debye length is 40 nm; for other dopings, L_D will vary as $1/\sqrt{N}$, that is, a reduction by a factor of 3.16 per decade.

2.2.2 Linearly Graded Junction

In practical devices, the doping profiles are not abrupt, especially near the metallurgical junction where the two types meet and they compensate each other. When the depletion widths terminate within this transition region, the doping profile can be approximated by a linear function. Consider the thermal-equilibrium case first. The impurity distribution for a linearly graded junction is shown in Fig. 5a. The Poisson equation for this case is

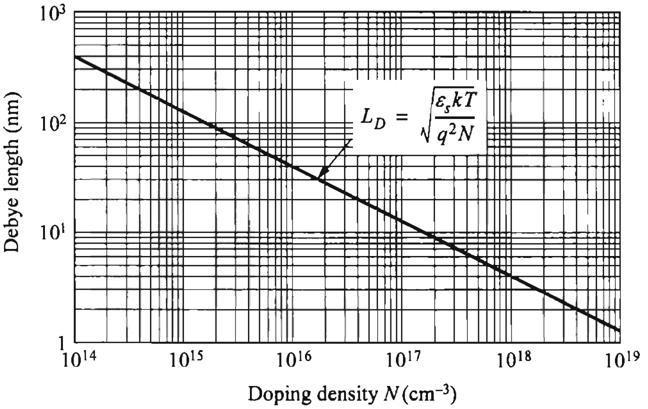


Fig. 4 Debye length in Si at room temperature as a function of doping density N.

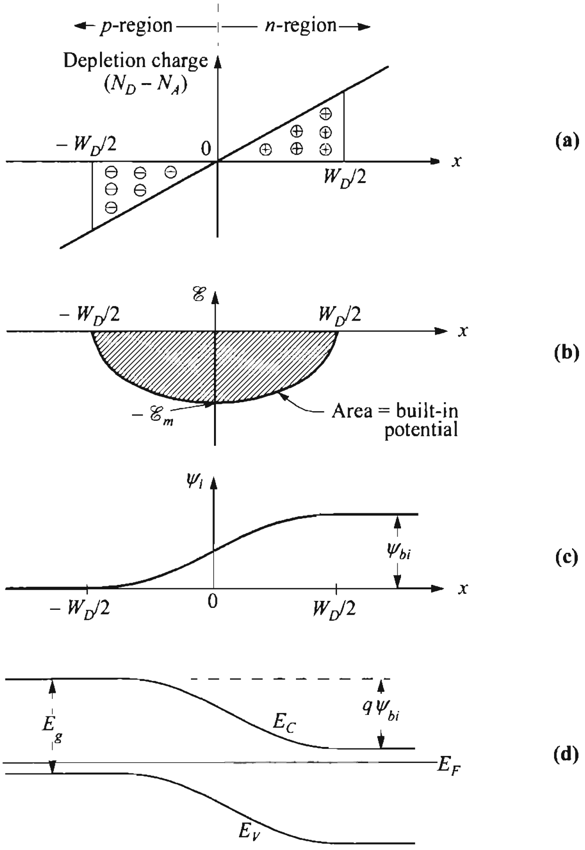

p-region
n-region
Depletion charge
$(N_D - N_A)$
$-W_D/2$
0
$W_D/2$
x
(a)
$\mathscr{E}$
$-W_D/2$
$W_D/2$
x
$-\mathscr{E}_m$
Area = built-in potential
(b)
ψ_i
ψ_{bi}
x
$-W_D/2$
0
$W_D/2$
(c)
E_g
E_C
$q\psi_{bi}$
E_F
E_V
(d)

Fig. 5 Linearly graded junction in thermal equilibrium. (a) Space-charge distribution. (b) Electric-field distribution. (c) Potential distribution. (d) Energy-band diagram.

$$-\frac{d^2\psi_i}{dx^2} = \frac{d\mathscr{E}}{dx} = \frac{\rho(x)}{\varepsilon_s} = \frac{q}{\varepsilon_s}(p - n + ax)$$

$$\approx \frac{qax}{\varepsilon_s} \qquad -\frac{W_D}{2} \le x \le \frac{W_D}{2} \tag{30}$$

where a is the doping gradient in cm^{-4}. By integrating Eq. 30 once, we obtain the field distribution shown in Fig. 5b:

$$\mathscr{E}(x) = -\frac{qa}{2\varepsilon_s}\left[\left(\frac{W_D}{2}\right)^2 - x^2\right] \qquad -\frac{W_D}{2} \le x \le \frac{W_D}{2} \tag{31}$$

with the maximum field $\mathscr{E}_m$ at $x = 0$,

$$|\mathscr{E}_m| = \frac{qaW_D^2}{8\varepsilon_s}. \tag{32}$$

Integrating Eq. 30 once again gives the potential distribution shown in Fig. 5c

$$\psi_i(x) = \frac{qa}{6\varepsilon_s}\left[2\left(\frac{W_D}{2}\right)^3 + 3\left(\frac{W_D}{2}\right)^2 x - x^3\right] \qquad -\frac{W_D}{2} \le x \le \frac{W_D}{2} \tag{33}$$

from which the built-in potential can be related to the depletion width

$$\psi_{bi} = \frac{qaW_D^3}{12\varepsilon_s} \tag{34}$$

or

$$W_D = \left(\frac{12\varepsilon_s\psi_{bi}}{qa}\right)^{1/3}. \tag{35}$$

Since the values of the impurity concentrations at the edges of the depletion region ($-W_D/2$ and $W_D/2$) are the same and equal to $aW_D/2$, the built-in potential for a linearly graded junction can be approximated by an expression similar to Eq. 5:

$$\begin{aligned}\psi_{bi} &\approx \frac{kT}{q}\ln\left[\frac{(aW_D/2)(aW_D/2)}{n_i^2}\right] \\ &\approx \frac{2kT}{q}\ln\left(\frac{aW_D}{2n_i}\right).\end{aligned} \tag{36}$$

Equations 35 and 36 can thus be used to solve for W_D and ψ_{bi}.

Based on an accurate numerical technique,[8] the built-in potential can be calculated explicitly by an expression as a *gradient voltage* V_g:

$$V_g = \frac{2kT}{3q}\ln\left(\frac{a^2\varepsilon_s kT}{8n_i^3 q^2}\right). \tag{37}$$

The gradient voltages for Si and GaAs as a function of impurity gradient are shown in Fig. 6. These voltages are smaller than the ψ_{bi} calculated from Eq. 36, using the depletion approximation, by more than 100 mV. The depletion-layer width and the corresponding capacitance for silicon using this V_g as the built-in potential are plotted in Fig. 7 as a function of net potential ($V_g - V$).

The depletion-layer capacitance for a linearly graded junction is given by

$$C_D = \frac{\varepsilon_s}{W_D} = \left[\frac{qa\varepsilon_s^2}{12(\psi_{bi} - V)}\right]^{1/3} \tag{38}$$

where V is positive/negative for forward/reverse bias.

2.2.3 Arbitrary Doping Profile

In this section we consider the doping near the junction to be of any arbitrary shape. Limiting the discussion to the n-side of a p^+-n junction, the net potential change at the junction is given by integrating the total field across the depletion region;

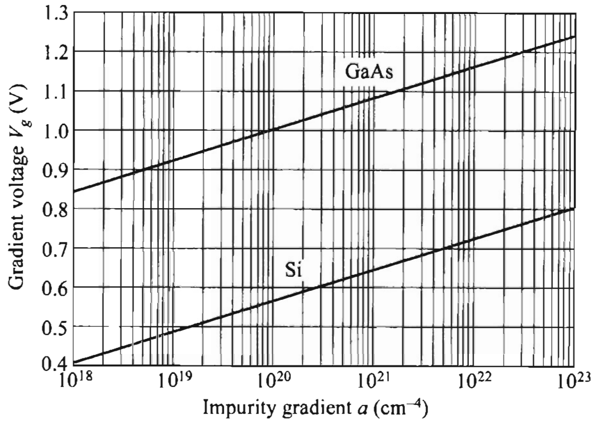


Fig. 6 Gradient voltages for linearly graded junctions in Si and GaAs.

$$\psi_n = \psi_{n0} - V = -\int_0^{W_D} \mathscr{E}(x)dx = -x\mathscr{E}(x)\Big|_0^{W_D} + \int_{\mathscr{E}(0)}^{\mathscr{E}(W_D)} x\,d\mathscr{E}, \tag{39}$$

where ψ_{n0} is ψ_n at zero bias. The first term becomes zero since the field at the depletion edge $\mathscr{E}(W_D)$ is zero. The interface potential becomes

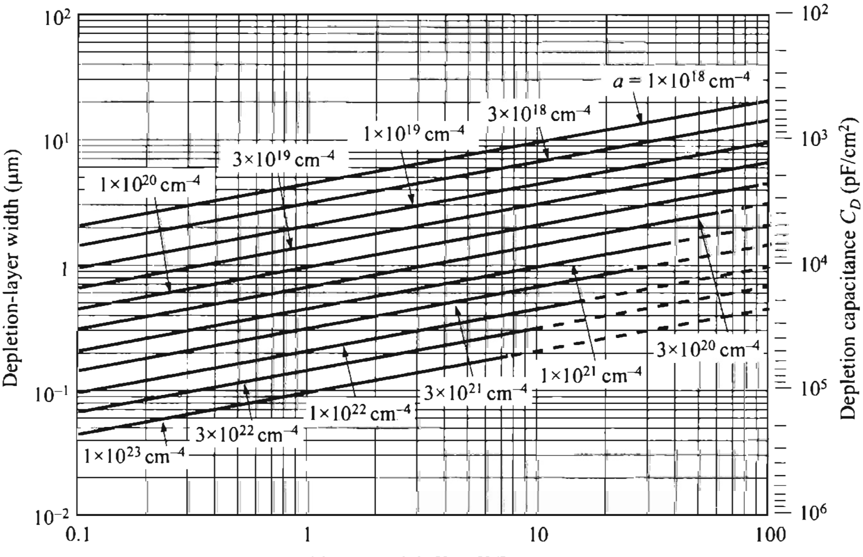


Fig. 7 Depletion-layer width and depletion-layer capacitance per unit area as a function of net potential ($V_g - V$) for different impurity gradients in linearly graded junctions in Si. Dashed lines represent breakdown conditions.

$$\psi_n = \int_{\mathscr{E}(0)}^{\mathscr{E}(W_D)} x\frac{d\mathscr{E}}{dx}dx = \frac{q}{\varepsilon_s}\int_0^{W_D} xN_D(x)dx. \tag{40}$$

Meanwhile, the total depletion-layer charge is given by

$$Q_D = q\int_0^{W_D} N_D(x)dx. \tag{41}$$

Differentiating the above quantities with respect to the depletion width gives

$$\frac{dV}{dW_D} = -\frac{d\psi_n}{dW_D} = -\frac{qN_D(W_D)W_D}{\varepsilon_s}, \tag{42}$$

$$\frac{dQ_D}{dW_D} = qN_D(W_D). \tag{43}$$

From these we obtain the depletion-layer capacitance,

$$C_D = \left|\frac{dQ_D}{dV}\right| = \left|\frac{dQ_D}{dW_D}\times\frac{dW_D}{dV}\right| = \frac{\varepsilon_s}{W_D}. \tag{44}$$

Again the general expression of ε_s/W_D is obtained and is applicable to any arbitrary doping profile. From this we can derive Eq. 26 for a general nonuniform profile;

$$\frac{d(1/C_D^2)}{dV} = \frac{d(1/C_D^2)dW_D}{dW_D\ dV} = \frac{2W_D dW_D}{\varepsilon_s^2\ dV}$$
$$= -\frac{2}{q\varepsilon_s N_D(W_D)}. \tag{45}$$

This C-V technique can be used to measure nonuniform doping profile. The $1/C_D^2$-V plot (like that shown in Fig. 3) would deviate from a straight line if the doping is not constant.

2.3 CURRENT-VOLTAGE CHARACTERISTICS

2.3.1 Ideal Case—Shockley Equation[1,2]

The ideal current-voltage characteristics are based on the following four assumptions: (1) the abrupt depletion-layer approximation; that is, the built-in potential and applied voltages are supported by a dipole layer with abrupt boundaries, and outside the boundaries the semiconductor is assumed to be neutral; (2) the Boltzmann approximation, similar to Eqs. 21 and 23 of Chapter 1, is valid; (3) the low-injection assumption; that is, the injected minority carrier densities are small compared with the majority-carrier densities; and (4) no generation-recombination current exists inside the depletion layer, and the electron and hole currents are constant throughout the depletion layer.

We first consider the Boltzmann relation. At thermal equilibrium this relation is given by

$$n = n_i \exp\left(\frac{E_F - E_i}{kT}\right), \tag{46a}$$

$$p = n_i \exp\left(\frac{E_i - E_F}{kT}\right). \tag{46b}$$

Obviously, at thermal equilibrium, the pn product from the above equations is equal to n_i^2. When voltage is applied, the minority-carrier densities on both sides of the junction are changed, and the pn product is no longer equal to n_i^2. We shall now define the quasi-Fermi (imref) levels as follows:

$$n \equiv n_i \exp\left(\frac{E_{Fn} - E_i}{kT}\right), \tag{47a}$$

$$p \equiv n_i \exp\left(\frac{E_i - E_{Fp}}{kT}\right), \tag{47b}$$

where E_{Fn} and E_{Fp} are the quasi-Fermi levels for electrons and holes, respectively. From Eqs. 47a and 47b we obtain

$$E_{Fn} \equiv E_i + kT\ln\left(\frac{n}{n_i}\right), \tag{48a}$$

$$E_{Fp} \equiv E_i - kT\ln\left(\frac{p}{n_i}\right). \tag{48b}$$

The pn product becomes

$$pn = n_i^2 \exp\left(\frac{E_{Fn} - E_{Fp}}{kT}\right). \tag{49}$$

For a forward bias, $(E_{Fn} - E_{Fp}) > 0$ and $pn > n_i^2$; on the other hand, for a reversed bias, $(E_{Fn} - E_{Fp}) < 0$ and $pn < n_i^2$.

From Eq. 156a of Chapter 1, Eq. 47a, and the fact that $\mathscr{E} \equiv \nabla E_i/q$, we obtain

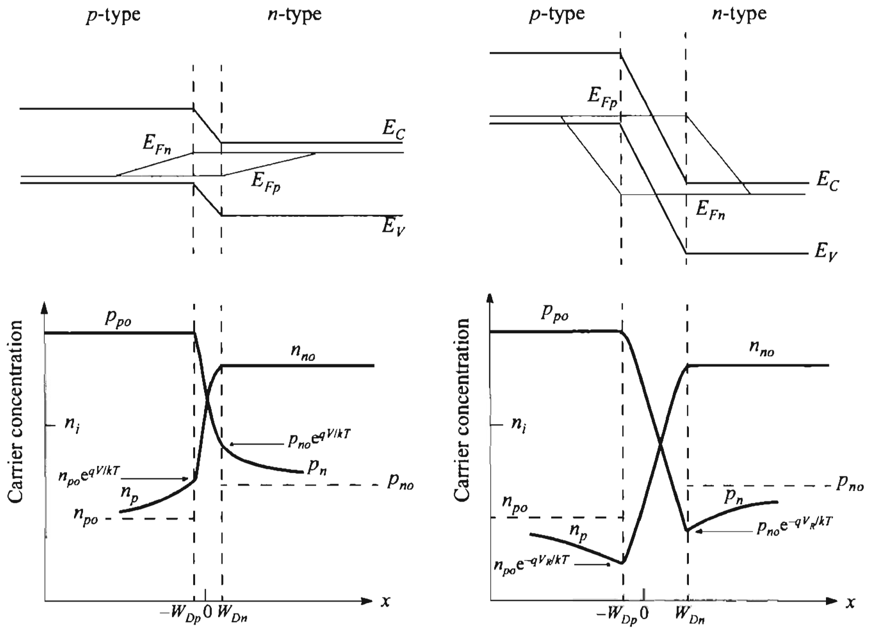
$$\begin{aligned} J_n &= q\mu_n\left(n\mathscr{E} + \frac{kT}{q}\nabla n\right) = \mu_n n \nabla E_i + \mu_n kT\left[\frac{n}{kT}(\nabla E_{Fn} - \nabla E_i)\right] \\ &= \mu_n n \nabla E_{Fn} . \end{aligned} \tag{50}$$

Similarly, we obtain,

$$J_p = \mu_p p \nabla E_{Fp}. \tag{51}$$

Thus, the electron and hole current densities are proportional to the gradients of the electron and hole quasi-Fermi levels, respectively. If $E_{Fn} = E_{Fp} =$ constant (at thermal equilibrium), then $J_n = J_p = 0$.

The idealized potential distributions and the carrier concentrations in a *p-n* junction under forward-bias and reverse-bias conditions are shown in Fig. 8. The variations of E_{Fn} and E_{Fp} with distance are related to the carrier concentrations as given in



(a) **(b)**

Fig. 8 Energy-band diagram, with quasi-Fermi levels for electrons and holes, and carrier distributions under (a) forward bias and (b) reverse bias.

Eqs. 48a and 48b, and to the current as given by Eqs. 50 and 51. Inside the depletion region, E_{Fn} and E_{Fp} remain relatively constant. This comes about because the carrier concentrations are relatively much higher inside the depletion region, but since the currents remain fairly constant, the gradients of the quasi-Fermi levels have to be small. In addition, the depletion width is typically much shorter than the diffusion length, so the total drop of quasi-Fermi levels inside the depletion width is not significant. With these arguments, it follows that within the depletion region,

$$qV = E_{Fn} - E_{Fp} . \qquad (52)$$

Equations 49 and 52 can be combined to give the electron density at the boundary of the depletion-layer region on the p-side ($x = -W_{Dp}$):

$$n_p(-W_{Dp}) = \frac{n_i^2}{p_p}\exp\left(\frac{qV}{kT}\right) \approx n_{po}\exp\left(\frac{qV}{kT}\right) \qquad (53a)$$

where $p_p \approx p_{po}$ for low-level injection, and n_{po} is the equilibrium electron density on the p-side. Similarly,

$$p_n(W_{Dn}) = p_{no}\exp\left(\frac{qV}{kT}\right) \qquad (53b)$$

at $x = W_{Dn}$ for the n-type boundary. The preceding equations are the most-important boundary conditions for the ideal current-voltage equation.

From the continuity equations we obtain for the steady-state condition in the n-side of the junction:

$$-U + \mu_n \mathscr{E} \frac{dn_n}{dx} + \mu_n n_n \frac{d\mathscr{E}}{dx} + D_n \frac{d^2 n_n}{dx^2} = 0 , \tag{54a}$$

$$-U - \mu_p \mathscr{E} \frac{dp_n}{dx} - \mu_p p_n \frac{d\mathscr{E}}{dx} + D_p \frac{d^2 p_n}{dx^2} = 0 . \tag{54b}$$

In these equations, U is the net recombination rate. Note that due to charge neutrality, majority carriers need to adjust their concentrations such that $(n_n - n_{no}) = (p_n - p_{no})$. It also follows that $dn_n/dx = dp_n/dx$. Multiplying Eq. 54a by $\mu_p p_n$ and Eq. 54b by $\mu_n n_n$, and combining with the Einstein relation $D = (kT/q)\mu$, we obtain

$$-\frac{p_n - p_{no}}{\tau_p} - \frac{n_n - p_n}{(n_n/\mu_p) + (p_n/\mu_n)} \frac{\mathscr{E} dp_n}{dx} + D_a \frac{d^2 p_n}{dx^2} = 0 \tag{55}$$

where

$$D_a = \frac{n_n + p_n}{n_n/D_p + p_n/D_n} \tag{56}$$

is the ambipolar diffusion coefficient, and

$$\tau_p \equiv \frac{p_n - p_{no}}{U} . \tag{57}$$

From the low-injection assumption [e.g., $p_n \ll (n_n \approx n_{no})$ in the n-type semiconductor], Eq. 55 reduces to

$$-\frac{p_n - p_{no}}{\tau_p} - \mu_p \mathscr{E} \frac{dp_n}{dx} + D_p \frac{d^2 p_n}{dx^2} = 0 \tag{58}$$

which is Eq. 54b except that the term $\mu_p p_n d\mathscr{E}/dx$ is ignored under the low-injection assumption.

In the neutral region where there is no electric field, Eq. 58 further reduces to

$$\frac{d^2 p_n}{dx^2} - \frac{p_n - p_{no}}{D_p \tau_p} = 0. \tag{59}$$

The solution of Eq. 59, with the boundary conditions of Eq. 53b and $p_n(x = \infty) = p_{no}$, gives

$$p_n(x) - p_{no} = p_{no}\left[\exp\left(\frac{qV}{kT}\right) - 1\right]\exp\left(-\frac{x - W_{Dn}}{L_p}\right) \tag{60}$$

where

$$L_p \equiv \sqrt{D_p \tau_p} . \tag{61}$$

At $x = W_{Dn}$, the hole diffusion current is

$$J_p = -qD_p\frac{dp_n}{dx}\bigg|_{W_{Dn}} = \frac{qD_p p_{no}}{L_p}\left[\exp\left(\frac{qV}{kT}\right) - 1\right]. \tag{62a}$$

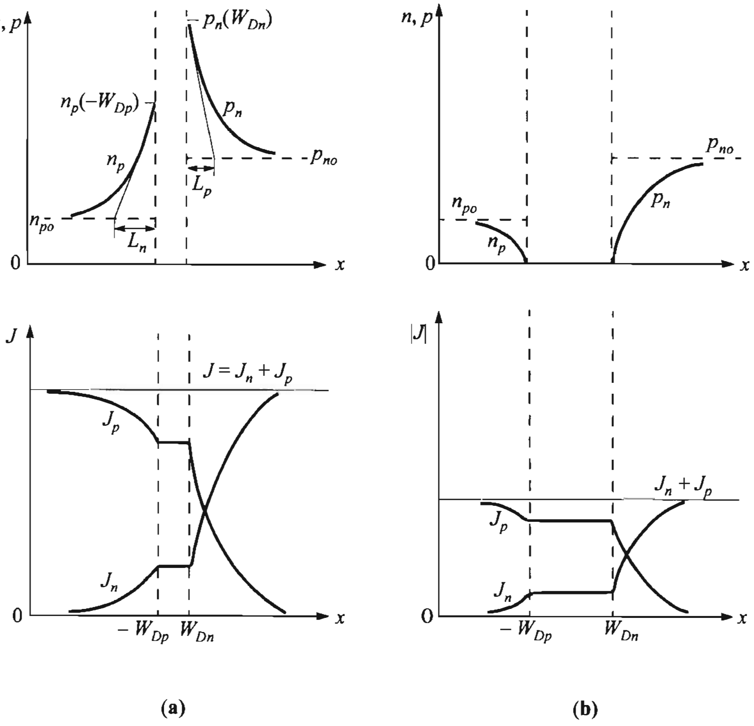
Similarly, we obtain the electron diffusion current in the p-side

$$J_n = qD_n\frac{dn_p}{dx}\bigg|_{-W_{Dp}} = \frac{qD_n n_{po}}{L_n}\left[\exp\left(\frac{qV}{kT}\right) - 1\right]. \tag{62b}$$

The minority-carrier densities and the current densities for the forward-bias and reverse-bias conditions are shown in Fig. 9. It is interesting to note that the hole current is due to injection of holes from the p-side to the n-side, but the magnitude is determined by the properties in the n-side only (D_p, L_p, p_{no}). The analogy holds for the electron current.

The total current is given by the sum of Eqs. 62a and 62b:

$$J = J_p + J_n = J_0\left[\exp\left(\frac{qV}{kT}\right) - 1\right], \tag{63}$$



(a) **(b)**

Fig. 9 Carrier distributions and current densities (both linear plots) for (a) forward-biased conditions and (b) reverse-biased conditions.

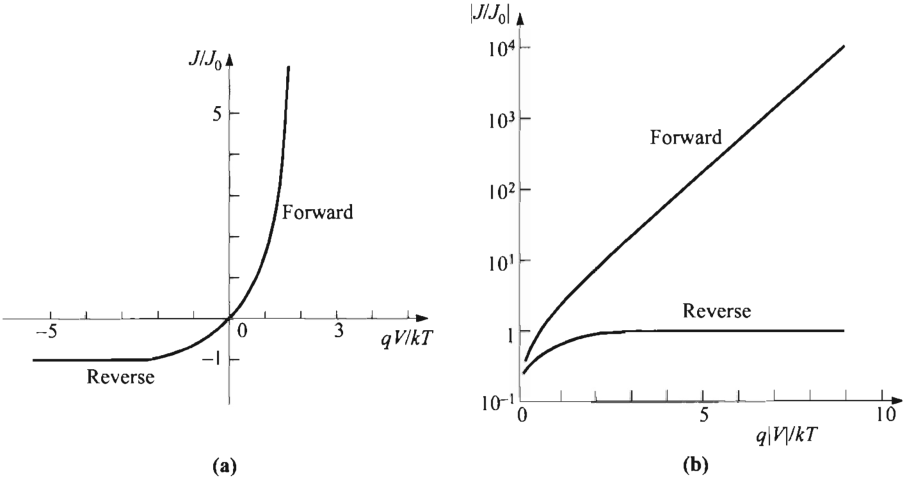
$$J_0 \equiv \frac{qD_p p_{no}}{L_p} + \frac{qD_n n_{po}}{L_n} \equiv \frac{qD_p n_i^2}{L_p N_D} + \frac{qD_n n_i^2}{L_n N_A} . \tag{64}$$

Equation 63 is the celebrated Shockley equation,[1,2] which is the ideal diode law. The ideal current-voltage relation is shown in Figs. 10a and b in the linear and semilog plots, respectively. In the forward direction (positive bias on the p-side) for $V > 3kT/q$, the rate of current rise is constant (Fig. 10b); at 300 K for every decade change of current, the voltage changes by 59.5 mV ($= 2.3kT/q$). In the reverse direction, the current density saturates at $-J_0$.

We shall now briefly consider the temperature effect on the saturation current density J_0. We shall consider only the first term in Eq. 64, since the second term will behave similarly to the first one. For the one-sided p^+-n abrupt junction (with donor concentration N_D), $p_{no} \gg n_{po}$, the second term can also be neglected. The quantities n_i, D_p, p_{no}, and L_p ($\equiv \sqrt{D_p \tau_p}$) are all temperature-dependent. If D_p/τ_p is proportional to T^γ, where γ is a constant, then

$$\begin{aligned} J_0 &\approx \frac{qD_p p_{no}}{L_p} \approx q\sqrt{\frac{D_p}{\tau_p}}\frac{n_i^2}{N_D} \propto T^{\gamma/2}\left[T^3 \exp\left(-\frac{E_g}{kT}\right)\right] \\ &\propto T^{(3+\gamma/2)} \exp\left(-\frac{E_g}{kT}\right) . \end{aligned} \tag{65}$$

The temperature dependence of the term $T^{(3+\gamma/2)}$ is not important compared with the exponential term. The slope of a plot J_0 versus $1/T$ is determined mainly by the energy gap E_g. It is expected that in the reverse direction, where $|J_R| \approx J_0$, the current will increase approximately as $\exp(-E_g/kT)$ with temperature; and in the forward direction, where $J_F \approx J_0 \exp(qV/kT)$, the current will increase approximately as $\exp[-(E_g - qV)/kT]$.



(a) (b)

Fig. 10 Ideal current-voltage characteristics. (a) Linear plot. (b) Semilog plot.

The Shockley equation adequately predicts the current-voltage characteristics of germanium *p-n* junctions at low current densities. For Si and GaAs *p-n* junctions, however, the ideal equation can only give qualitative agreement. The departures from the ideal are mainly due to: (1) the generation and recombination of carriers in the depletion layer, (2) the high-injection condition that may occur even at relatively small forward bias, (3) the parasitic *IR* drop due to series resistance, (4) the tunneling of carriers between states in the bandgap, and (5) the surface effects. In addition, under sufficiently larger field in the reverse direction, the junction will breakdown as a result, for example, of avalanche multiplication. The junction breakdown will be discussed in Section 2.4.

The surface effects on *p-n* junctions are primarily due to ionic charges on or outside the semiconductor surface that induce image charges in the semiconductor, and thereby cause the formation of the so-called surface channels or surface depletion-layer regions. Once a channel is formed, it modifies the junction depletion region and gives rise to surface leakage current. For Si planar *p-n* junctions, the surface leakage current is generally much smaller than the generation-recombination current in the depletion region.

2.3.2 Generation-Recombination Process[3]

Consider first the generation current under the reverse-bias condition. Because of the reduction in carrier concentration under reverse bias ($pn \ll n_i^2$), the dominant generation processes, as discussed in Section 1.5.4, are those of emission. The rate of generation of electron-hole pairs can be obtained from Eq. 92 of Chapter 1 with the condition $p \ll n_i$ and $n \ll n_i$:

$$U = -\left\{\frac{\sigma_p \sigma_n v_{th} N_t}{\sigma_n \exp[(E_t - E_i)/kT] + \sigma_p \exp[(E_i - E_t)/kT]}\right\} n_i \equiv -\frac{n_i}{\tau_g} \tag{66}$$

where τ_g is the generation lifetime and is defined as the reciprocal of the expression in brackets (see Eq. 98 of Chapter 1 and the discussion following). The current due to generation in the depletion region is thus given by

$$J_{ge} = \int_0^{W_D} q|U| dx \approx q|U| W_D \approx \frac{q n_i W_D}{\tau_g} \tag{67}$$

where W_D is the depletion-layer width. If the generation lifetime is a slowly varying function of temperature, the generation current will then have the same temperature dependence as n_i. At a given temperature, J_{ge} is proportional to the depletion-layer width, which in turn is dependent on the applied reverse bias. It is thus expected that

$$J_{ge} \propto (\psi_{bi} + V)^{1/2} \tag{68}$$

for abrupt junctions, and

$$J_{ge} \propto (\psi_{bi} + V)^{1/3} \tag{69}$$

for linearly graded junctions.

The total reverse current (for $p_{no} \gg n_{po}$ and $|V| > 3kT/q$) can be approximated by the sum of the diffusion component in the neutral region and the generation current in the depletion region:

$$J_R = q\sqrt{\frac{D_p}{\tau_p}}\frac{n_i^2}{N_D} + \frac{qn_iW_D}{\tau_g}. \tag{70}$$

For semiconductors with large values of n_i (such as Ge), the diffusion component will dominate at room temperature and the reverse current will follow the Shockley equation; but if n_i is small (such as for Si), the generation current may dominate. A typical result for Si is shown in Fig. 11, curve (e). At sufficiently high temperatures, however, the diffusion current will dominate.

At forward bias, where the major recombination-generation processes in the depletion region are the capture processes, we have a recombination current J_{re} in addition to the diffusion current. Substituting Eq. 49 in Eq. 92 of Chapter 1 yields

$$U = \frac{\sigma_p\sigma_n v_{th}N_t n_i^2[\exp(qV/kT)-1]}{\sigma_n\{n+n_i\exp[(E_t-E_i)/kT]\}+\sigma_p\{p+n_i\exp[(E_i-E_t)/kT]\}}. \tag{71}$$

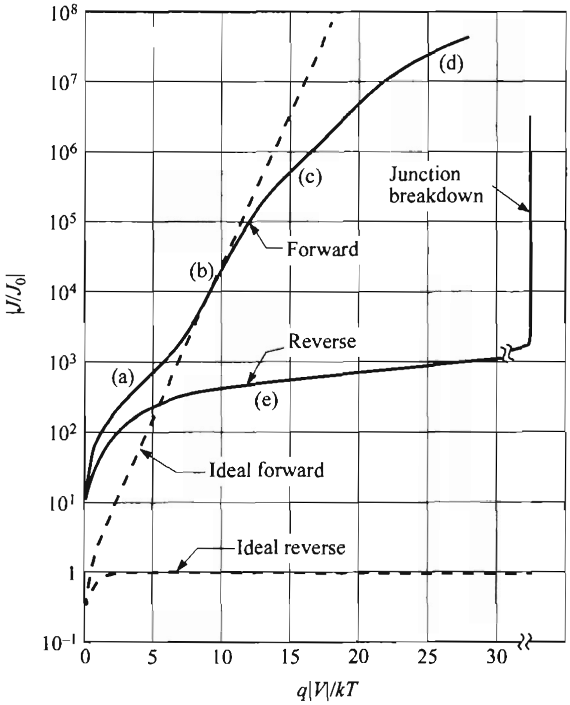


Fig. 11 Current-voltage characteristics of a practical Si diode. (a) Generation-recombination current region. (b) Diffusion-current region. (c) High-injection region. (d) Series-resistance effect. (e) Reverse leakage current due to generation-recombination and surface effects.

Under the assumptions that $E_t = E_i$ and $\sigma_n = \sigma_p = \sigma$, Eq. 71 reduces to

$$U = \frac{\sigma v_{th} N_t n_i^2 [\exp(qV/kT) - 1]}{n + p + 2n_i}$$
$$= \frac{\sigma v_{th} N_t n_i^2 [\exp(qV/kT) - 1]}{n_i \{\exp[(E_{Fn} - E_i)/kT] + \exp[(E_i - E_{Fp})/kT] + 2\}}. \tag{72}$$

The maximum value of U exists in the depletion region where E_i is halfway between E_{Fn} and E_{Fp}, and so the denominator of Eq. 72 becomes $2n_i[\exp(qV/2kT) + 1]$. We obtain for $V > kT/q$,

$$U \approx \frac{1}{2} \sigma v_{th} N_t n_i \exp\left(\frac{qV}{2kT}\right) \tag{73}$$

and

$$J_{re} = \int_0^{W_D} qU dx \approx \frac{qW_D}{2} \sigma v_{th} N_t n_i \exp\left(\frac{qV}{2kT}\right) \approx \frac{qW_D n_i}{2\tau} \exp\left(\frac{qV}{2kT}\right). \tag{74}$$

The above approximation assumes that most part of the depletion layer has this maximum recombination rate, and J_{re} is thus somewhat an overestimate. A more rigorous derivation gives[9]

$$J_{re} = \int_0^{W_D} qU dx = \sqrt{\frac{\pi}{2}} \frac{kTn_i}{\tau \mathscr{E}_o} \exp\left(\frac{qV}{2kT}\right) \tag{75}$$

where $\mathscr{E}_o$ is the electric field at the location of maximum recombination, and it is equal to

$$\mathscr{E}_o = \sqrt{\frac{qN(2\psi_B - V)}{\varepsilon_s}}. \tag{76}$$

Similar to the generation current in reverse bias, the recombination current in forward bias is also proportional to n_i. The total forward current can be approximated by the sum of Eqs. 63 and 75. For a p^+-n junction ($p_{no} \gg n_{po}$) and $V \gg kT/q$:

$$J_F = q\sqrt{\frac{D_p}{\tau_p}} \frac{n_i^2}{N_D} \exp\left(\frac{qV}{kT}\right) + \sqrt{\frac{\pi}{2}} \frac{kTn_i}{\tau_p \mathscr{E}_o} \exp\left(\frac{qV}{2kT}\right). \tag{77}$$

The experimental results in general can be represented by the empirical form,

$$J_F \propto \exp\left(\frac{qV}{\eta kT}\right) \tag{78}$$

where the ideality factor η equals 2 when the recombination current dominates [Fig. 11, curve (a)] and η equals 1 when the diffusion current dominates [Fig. 11, curve (b)]. When both currents are comparable, η has a value between 1 and 2.

2.3.3 High-Injection Condition

At high current densities (under the forward-bias condition) such that the injected minority-carrier density is comparable to the majority concentration, both drift and diffusion current components must be considered. The individual conduction current densities can always be given by Eqs. 50 and 51. Since J_p, q, μ_p, and p are positive, the quasi-Fermi level for holes E_{Fp} increases monotonically to the right as shown in Fig. 8a. Similarly, the quasi-Fermi level for electrons E_{Fn} decreases monotonically to the left. Thus, everywhere the separation of the two quasi-Fermi levels must be equal to or less than the applied voltage, and therefore[10]

$$pn \le n_i^2 \exp\left(\frac{qV}{kT}\right) \tag{79}$$

even under the high-injection condition. Note also that the foregoing argument does not depend on recombination in the depletion region.

To illustrate the high-injection case, we present in Fig. 12 plots of numerical simulation results for carrier concentrations and energy-band diagram with quasi-Fermi levels for a silicon p^+-n step junction. The current densities in Figs. 12a, b, and c are 10, 10^3, and 10^4 A/cm^2, respectively. At 10 A/cm^2 the diode is in the low-injection regime. Almost all of the potential drop occurs across the junction. The hole concentration in the n-side is small compared to the electron concentration. At 10^3 A/cm^2 the electron concentration near the junction exceeds the donor concentration appreciably (bear in mind that from charge neutrality, injected carriers $\Delta p = \Delta n$). An ohmic potential drop appears on the n-side. At 10^4 A/cm^2 we have very high injection; the poten-

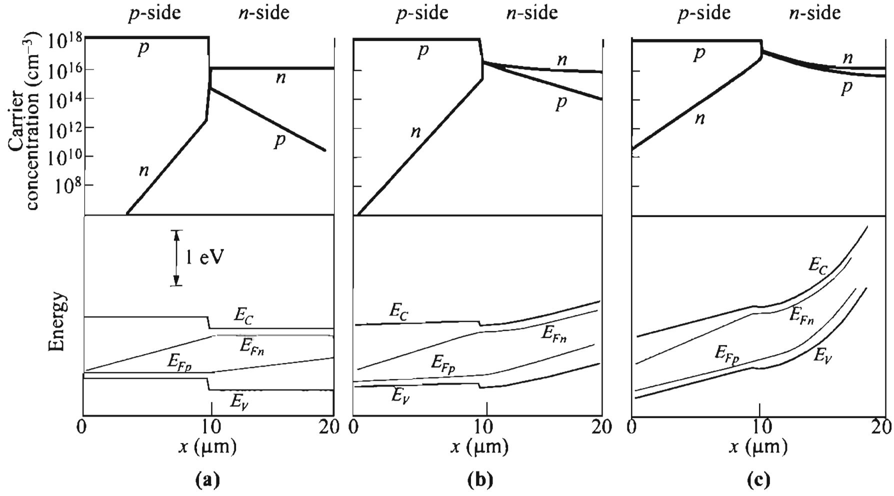


Fig. 12 Carrier concentrations and energy-band diagrams for a Si p^+-n junction operated at different current densities. (a) 10 A/cm^2. (b) 10^3 A/cm^2. (c) 10^4 A/cm^2. Device parameters: $N_A = 10^{18}$ cm^{-3}, $N_D = 10^{16}$ cm^{-3}, $\tau_n = 3\times10^{-10}$ s, and $\tau_p = 8.4\times10^{-10}$ s. (After Ref. 10.)

tial drop across the junction is insignificant compared to ohmic drops on both sides of the neutral regions. Even though only the center region of the diode is shown in Fig. 12, it is apparent that the separation of the quasi-Fermi levels is equal to or less than the applied voltage (qV).

From Fig. 12b and c, the carrier densities at the n-side of the junction are comparable ($n = p$). Substituting this condition in Eq. 79, we obtain $p_n(x = W_{Dn}) \approx n_i \exp(qV/2kT)$. The current then becomes roughly proportional to $\exp(qV/2kT)$, as shown in Fig. 11, curve (c).

At high-current levels we should consider another effect associated with the finite resistivity in the quasi-neutral regions. This resistance absorbs an appreciable amount of the applied voltage between the diode terminals. This is shown in Fig. 11 as curve-(d). One can estimate the series resistance from comparing the experimental curve to the ideal curve ($\Delta V = IR$). The series resistance effect can be substantially reduced by the use of epitaxial materials (p^+-n-n^+).

2.3.4 Diffusion Capacitance

The depletion-layer capacitance considered previously accounts for most of the junction capacitance when the junction is reverse-biased. When forward-biased, there is, in addition, a significant contribution to junction capacitance from the rearrangement of minority carrier density, the so-called diffusion capacitance. In other words, the latter is due to the injected charge, while the former to the depletion-layer charge.

When a small ac signal is applied to a junction that is forward-biased at a dc voltage V_0 and current density J_0, the total voltage and current are defined by

$$V(t) = V_0 + V_1 \exp(j\omega t), \tag{80}$$

$$J(t) = J_0 + J_1 \exp(j\omega t) \tag{81}$$

where V_1 and J_1 are the small-signal voltage and current density, respectively. The imaginary part of the admittance J_1/V_1 will give the diffusion conductance and diffusion capacitance:

$$Y \equiv \frac{J_1}{V_1} \equiv G_d + j\omega C_d\ . \tag{82}$$

The electron and hole densities at the depletion region boundaries can be obtained from Eqs. 53a and 53b by using $[V_0 + V_1\exp(j\omega t)]$ instead of V. We obtain for the n-side of the junction and $V_1 \ll V_0$,

$$p_n(W_{Dn}) = p_{no}\exp\left\{\frac{q[V_0 + V_1\exp(j\omega t)]}{kT}\right\}$$
$$\approx p_{no}\exp\left(\frac{qV_0}{kT}\right) + \frac{p_{no}qV_1}{kT}\exp\left(\frac{qV_0}{kT}\right)\exp(j\omega t) \approx p_{no}\exp\left(\frac{qV_0}{kT}\right) + \tilde{p}_n(t)\ . \tag{83}$$

A similar expression can be obtained for the electron density in the p-side. The first term in Eq. 83 is the dc component, and the second term is the small-signal ac component. Substituting $\tilde{p}_n$ into the continuity equation (Eq. 158b of Chapter 1 with $G_p = \mathscr{E} = d\mathscr{E}/dx = 0$) yields

$$j\omega\tilde{p}_n = -\frac{\tilde{p}_n}{\tau_p} + D_p\frac{d^2\tilde{p}_n}{dx^2} \tag{84}$$

or

$$\frac{d^2\tilde{p}_n}{dx^2} - \frac{\tilde{p}_n}{D_p\tau_p/(1+j\omega\tau_p)} = 0. \tag{85}$$

Equation 85 is identical to Eq. 59 if the carrier lifetime is expressed as

$$\tau_p^* = \frac{\tau_p}{1+j\omega\tau_p}. \tag{86}$$

We can then obtain the alternating current density from Eq. 63 by making the appropriate substitutions:

$$\begin{aligned} J &= \left(qp_{no}\sqrt{\frac{D_p}{\tau_p^*}} + qn_{po}\sqrt{\frac{D_n}{\tau_n^*}}\right)\exp\left\{\frac{q[V_0 + V_1\exp(j\omega t)]}{kT}\right\} \\ &\approx \left(qp_{no}\sqrt{\frac{D_p}{\tau_p^*}} + qn_{po}\sqrt{\frac{D_n}{\tau_n^*}}\right)\left[\exp\left(\frac{qV_0}{kT}\right)\right]\left[1 + \frac{qV_1}{kT}\exp(j\omega t)\right], \end{aligned} \tag{87}$$

with the ac component being

$$J_1 = \left(\frac{qD_pp_{no}\sqrt{1+j\omega\tau_p}}{L_p} + \frac{qD_nn_{po}\sqrt{1+j\omega\tau_n}}{L_n}\right)\left[\exp\left(\frac{qV_0}{kT}\right)\right]\frac{qV_1}{kT}. \tag{88}$$

From J_1/V_1, both G_d and C_d can be found and they are frequency dependent.

For relatively low frequencies ($\omega\tau_p$, $\omega\tau_n \ll 1$), the diffusion conductance G_{d0} is given by

$$G_{d0} = \frac{q}{kT}\left(\frac{qD_pp_{no}}{L_p} + \frac{qD_nn_{po}}{L_n}\right)\exp\left(\frac{qV_0}{kT}\right) \qquad \text{mho/cm}^2 \tag{89}$$

which has exactly the same value obtained by differentiating Eq. 63. The low-frequency diffusion capacitance C_{d0} can be obtained by using the approximation $\sqrt{1+j\omega\tau} \approx (1 + 0.5j\omega\tau)$

$$C_{d0} = \frac{q^2}{2kT}(L_pp_{no} + L_nn_{po})\exp\left(\frac{qV_0}{kT}\right) \qquad \text{F/cm}^2. \tag{90}$$

This diffusion capacitance is proportional to the forward current. For an n^+-p one-sided junction, it can shown that

$$C_{d0} = \frac{qL_n^2}{2kTD_n}J_F. \tag{91}$$

The frequency dependence of the diffusion conductance and capacitance is shown in Fig. 13 as a function of the normalized frequency $\omega\tau$ where only one term in Eq. 88

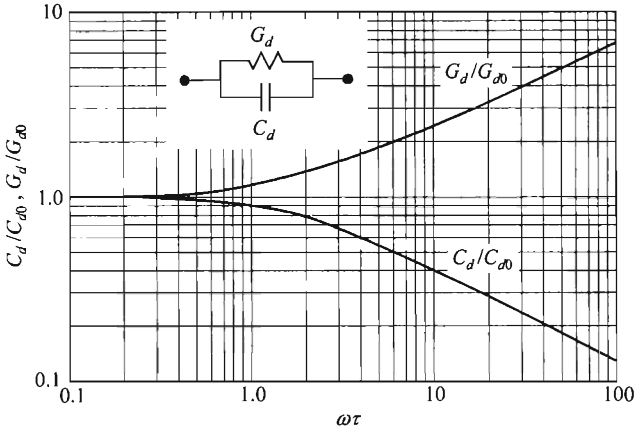


Fig. 13 Normalized diffusion conductance and diffusion capacitance versus $\omega\tau$. Inset shows the equivalent circuit of a p-n junction under forward bias.

is considered (e.g., the term contains p_{no} if $p_{no} \gg n_{po}$). The inset shows the equivalent circuit of the ac admittance. It is clear from Fig. 13 that the diffusion capacitance decreases with increasing frequency. For high frequencies, C_d is approximately proportional to $\omega^{-1/2}$. The diffusion capacitance is also proportional to the dc current level [$\propto \exp(qV_0/kT)$]. For this reason, C_d is especially important at low frequencies and under forward-bias conditions.

2.4 JUNCTION BREAKDOWN

When a sufficiently high field is applied to a *p-n* junction, the junction *breaks down* and conducts a very large current.[11] Breakdown occurs only in the reverse-bias regime because high voltage can be applied resulting in high field. There are basically three breakdown mechanisms: (1) thermal instability, (2) tunneling, and (3) avalanche multiplication. We consider the first two mechanisms briefly, and discuss avalanche multiplication in more detail.

2.4.1 Thermal Instability

Breakdown due to thermal instability is responsible for the maximum dielectric strength in most insulators at room temperature, and is also a major effect in semiconductors with relatively small bandgaps (e.g., Ge). Because of the heat dissipation caused by the reverse current at high reverse voltage, the junction temperature increases. This temperature increase, in turn, increases the reverse current in comparison with its value at lower voltages. This positive feedback is responsible for breakdown. The temperature effect on the reverse current-voltage characteristics is explained in Fig. 14. In this figure the reverse currents J_0 are represented by a family

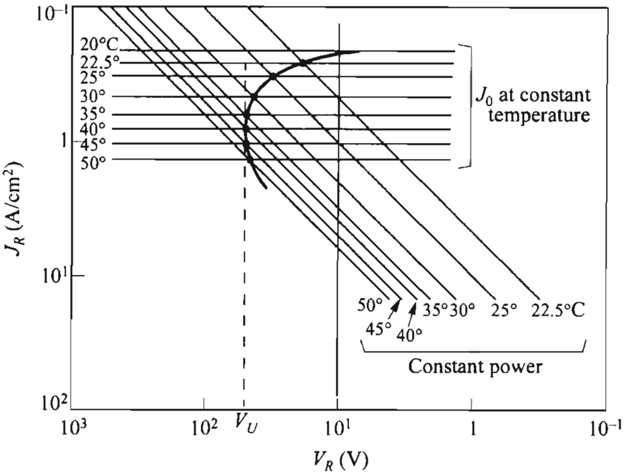


Fig. 14 Reverse current-voltage characteristics of thermal breakdown, where V_U is the turnover voltage. (Note decreasing values of coordinates.) (After Ref. 12.)

of horizontal lines. Each line represents the current at a constant junction temperature, and the current varies as $T^{3+\gamma/2}\exp(-E_g/kT)$, as discussed previously. The heat dissipation hyperbolas which are proportional to the power, given by the I-V product, are shown as sloped straight lines in the log-log plot. These lines also have to satisfy the curves of constant junction temperature. So the reverse current-voltage characteristics are obtained by the intersection points of these two sets of curves. Because of the heat dissipation at high reverse voltage, the characteristics show a negative differential resistance. In this condition, the diode will be destroyed unless some special measure such as a large series-limiting resistor is used. This effect is called thermal instability or thermal runaway. The voltage V_U is called the turnover voltage. For *p-n* junctions with relatively large saturation currents (e.g., in Ge), the thermal instability is important at room temperature, but at very low temperatures it becomes less important compared with other mechanisms.

2.4.2 Tunneling

We next consider the tunneling effect (see Section 1.5.7) when the junction is under a large reverse bias. It is well known that carriers can tunnel through a potential barrier if this barrier is sufficiently thin, induced by a large field as shown in Fig. 15a. In this particular case, the barrier has a triangular shape with the maximum height given by the energy gap. The derivation of the tunneling current of a *p-n* junction (tunnel diode) is considered in details in Chapter 8, and the result is given here as:

$$J_t = \frac{\sqrt{2m^*}q^3\mathscr{E}V_R}{4\pi^2\hbar^2\sqrt{E_g}}\exp\left(-\frac{4\sqrt{2m^*}E_g^{3/2}}{3q\mathscr{E}\hbar}\right). \tag{92}$$

Since the field is not constant, $\mathscr{E}$ is some average field inside the junction.

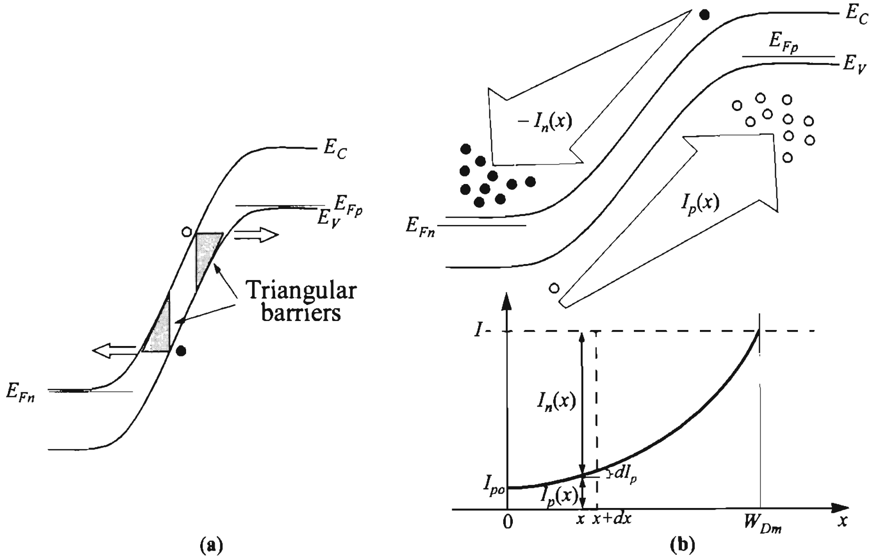


Fig. 15 Energy band diagrams showing breakdown mechanisms of (a) tunneling and (b) avalanche multiplication (example initiated by hole current I_{po}).

When the field approaches 10^6 V/cm in Si, significant current begins to flow by means of this band-to-band tunneling process. To obtain such a high field, the junction must have relatively high impurity concentrations on both the *p*- and *n*-side. The mechanism of breakdown for *p-n* junctions with breakdown voltages less than about $4E_g/q$ is due to the tunneling effect. For junctions with breakdown voltages in excess of $6E_g/q$, the mechanism is caused by avalanche multiplication. At voltages between 4 and 6 E_g/q, the breakdown is due to a mixture of both avalanche and tunneling. Since the energy bandgaps E_g in Si and GaAs decrease with increasing temperature (refer to Chapter 1), the breakdown voltage in these semiconductors due to the tunneling effect has a negative temperature coefficient; that is, the breakdown voltage decreases with increasing temperature. This is because a given breakdown current J_t can be reached at smaller reverse voltages (or fields) at higher temperatures (Eq. 92). This temperature effect is generally used to distinguish the tunneling mechanism from the avalanche mechanism, which has a positive temperature coefficient; that is, the breakdown voltage increases with increasing temperature.

2.4.3 Avalanche Multiplication

Avalanche multiplication, or impact ionization, is the most-important mechanism in junction breakdown. The avalanche breakdown voltage imposes an upper limit on the reverse bias for most diodes, on the collector voltage of bipolar transistors, and on the drain voltages of MESFETs and MOSFETs. In addition, the impact ionization mechanism can be used to generate microwave power, as in IMPATT devices, and to amplify optical signals, as in avalanche photodetectors.

We first derive the basic ionization integral which determines the breakdown condition. Assume that a current I_{po} is incident at the left-hand side of the depletion region with width W_{Dm} (Fig. 15b). If the electric field in the depletion region is high enough that electron-hole pairs are generated by the impact ionization process, the hole current I_p will increase with distance through the depletion region and reach a value M_pI_{po} at $x = W_{Dm}$. Similarly, the electron current I_n will increase from $I_n(W_{Dm}) = 0$ to $I_n(0) = I - I_{po}$, where the total current I $(= I_p + I_n)$ is constant at steady state. The incremental hole current is equal to the number of electron-hole pairs generated per second in the distance dx,

$$dI_p = I_p\alpha_p dx + I_n\alpha_n dx \tag{93}$$

or

$$\frac{dI_p}{dx} - (\alpha_p - \alpha_n)I_p = \alpha_n I. \tag{94}$$

The electron and hole ionization rates (α_n and α_p) have been considered in Chapter 1.

The solution of Eq. 94 with the boundary condition of $I = I_p(W_{Dm}) = M_pI_{po}$ is given by*

$$I_p(x) = I\left\{\int_0^x \alpha_n \exp\left[-\int_0^x (\alpha_p - \alpha_n)dx'\right]dx + \frac{1}{M_p}\right\}\Big/\exp\left[-\int_0^x (\alpha_p - \alpha_n)dx'\right] \tag{95}$$

where M_p is the multiplication factor of holes and is defined as

$$M_p \equiv \frac{I_p(W_{Dm})}{I_p(0)} \equiv \frac{I}{I_{po}}. \tag{96}$$

With a relationship†

$$\int_0^{W_{Dm}} (\alpha_p - \alpha_n)\exp\left[-\int_0^x (\alpha_p - \alpha_n)dx'\right]dx = -\exp\left[-\int_0^x (\alpha_p - \alpha_n)dx'\right]\Bigg|_0^{W_{Dm}}$$

$$= -\exp\left(\left[-\int_0^{W_{Dm}} (\alpha_p - \alpha_n)dx'\right] + 1\right), \tag{97}$$

Equation 95 can be evaluated at $x = W_{Dm}$ and be rewritten as

---

* Equation 94 has the form $y' + Py = Q$, where $y = I_p$. The standard solution is

$$y = \left[\int_0^x Q\left(\exp\int_0^x Pdx'\right)dx + C\right]\Big/\exp\int_0^x Pdx'$$

where C is the constant of integration.

$$1-\frac{1}{M_p} = \int_0^{W_{Dm}} \alpha_p \exp\left[-\int_0^x (\alpha_p - \alpha_n)dx'\right]dx. \tag{98}$$

Note that M_p is a function of α_n in addition to α_p. The avalanche breakdown voltage is defined as the voltage where M_p approaches infinity. Hence the breakdown condition is given by the ionization integral

$$\int_0^{W_{Dm}} \alpha_p \exp\left[-\int_0^x (\alpha_p - \alpha_n)dx'\right]dx = 1. \tag{99a}$$

If the avalanche process is initiated by electrons instead of holes, the ionization integral is given by

$$\int_0^{W_{Dm}} \alpha_n \exp\left[-\int_x^{W_{Dm}} (\alpha_n - \alpha_p)dx'\right]dx = 1. \tag{99b}$$

Equations 99a and 99b are equivalent;[13] that is, the breakdown condition depends only on what is happening within the depletion region and not on the carriers (or primary current) that initiate the avalanche process. The situation does not change when a mixed primary current initiates the breakdown, so either Eq. 99a or Eq. 99b gives the breakdown condition. For semiconductors with equal ionization rates ($\alpha_n = \alpha_p = \alpha$) such as GaP, Eq. 99a or 99b reduces to the simple expression

$$\int_0^{W_{Dm}} \alpha dx = 1. \tag{100}$$

From the breakdown conditions described above and the field dependence of the ionization rates, the breakdown voltage, maximum electric field, and depletion-layer width can be calculated. As discussed previously, the electric field and potential in the depletion layer are determined from the solutions of the Poisson equation. Depletion-layer boundaries that satisfy Eq. 99a or 99b can be obtained numerically using an iteration method. With known boundaries we obtain the breakdown voltage

$$V_{BD} = \frac{\mathscr{E}_m W_{Dm}}{2} = \frac{\varepsilon_s \mathscr{E}_m^2}{2qN} \tag{101}$$

for one-sided abrupt junctions, and

---

† Let

$$U = \int_0^x y dx' , \quad \frac{dU}{dx} = y , \quad \frac{d}{dU}e^U = e^U .$$

The integral can be simplified to

$$\int y\left(\exp\int_0^x y dx'\right)dx = \int y e^U dx = \int e^U dU = e^U = \exp\int_0^x y dx' .$$

$$V_{BD} = \frac{2\mathscr{E}_m W_{Dm}}{3} = \frac{4\mathscr{E}_m^{3/2}}{3}\left(\frac{2\varepsilon_s}{qa}\right)^{1/2} \tag{102}$$

for linearly graded junctions, where N is the ionized background impurity concentration of the lightly doped side, a the impurity gradient, and $\mathscr{E}_m$ the maximum field.

Figure 16a shows the calculated breakdown voltage as a function of N for abrupt junctions in Si, ⟨100⟩-oriented GaAs, and GaP. The experimental results are generally in good agreement with the calculated values.[15] The dashed line in the figure indicates the upper limit of N for which the avalanche breakdown calculation is valid. This limitation is based on the criterion of $6E_g/q$. Above these corresponding values of N, the tunneling mechanism will contribute to the breakdown process and eventually dominates.

In GaAs, the ionization rates and thus breakdown voltage depend on crystal orientations, besides doping concentration (refer to Chapter 1).[16] At a doping concentration of around 10^{16} cm^{-3}, the breakdown voltages are essentially independent of orientations. At lower dopings, V_{BD} in ⟨111⟩ becomes the largest whereas at higher dopings, V_{BD} in ⟨100⟩ is the largest.

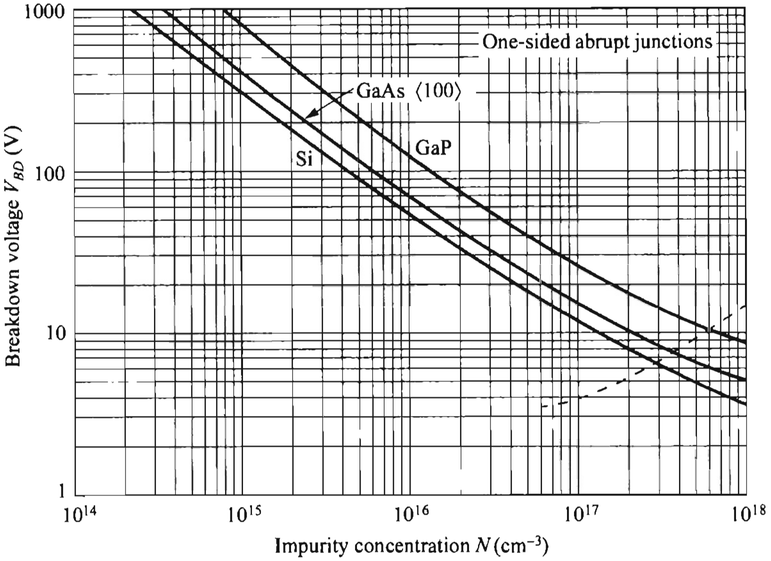
Figure 16b shows the calculated breakdown voltage versus the impurity gradient for linearly graded junctions. The dashed line indicates the upper limit of a for which the avalanche breakdown calculation is valid.

The calculated values of the maximum field $\mathscr{E}_m$ and the depletion-layer width at breakdown for the three semiconductors above are shown in Fig. 17a for the abrupt junctions, and in Fig. 17b for the linearly graded junctions. For the Si abrupt junctions, the maximum field at breakdown can be expressed as[17]

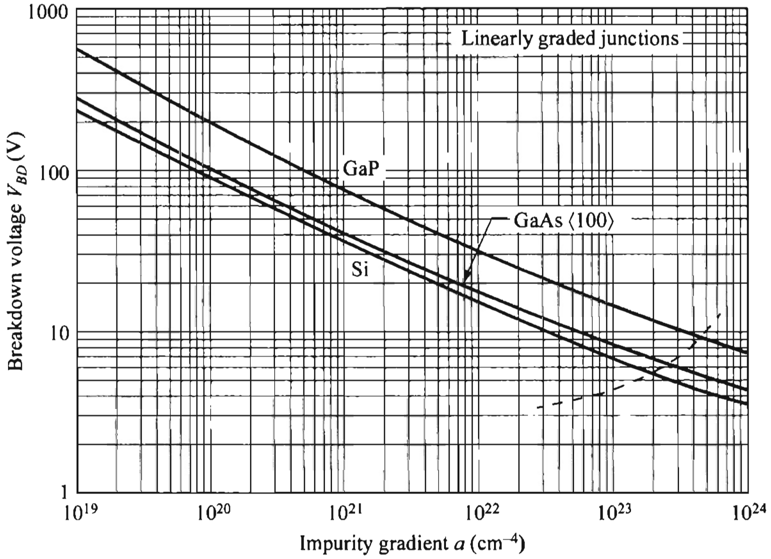
$$\mathscr{E}_m = \frac{4\times10^5}{1-(1/3)\log_{10}(N/10^{16}\ \text{cm}^{-3})} \quad \text{V/cm} \tag{103}$$

where N is in cm^{-3}.

Because of the strong dependence of the ionization rates on the field, the maximum field at breakdown, sometimes called the *critical field*, varies very slowly with either N or a (within a factor of 4 over many orders of magnitude in N and a). Thus, as a first approximation, we can assume that for a given semiconductor, $\mathscr{E}_m$ has a fixed value. Then from Eqs. 101 and 102 we obtain $V_{BD} \propto N^{1.0}$ for abrupt junctions and $V_{BD} \propto a^{-0.5}$ for linearly graded junctions. Figure 16 shows that the foregoing patterns are generally followed (within a factor of 3). Also as expected, for a given N or a, the breakdown voltage increases with the energy bandgap of the material, since the avalanche process requires band-to-band excitations. It should be cautioned that the critical field is only a rough guide line but not a fundamental material property. It assumes a uniform field over a large distance. For example, if there is a high field but only occurring over a small distance, breakdown would not happen since Eq. 100 cannot be satisfied. Also, the total voltage (field times distance) needs to be larger than the bandgap for band-to-band carrier multiplication. An example is the high field but small voltage drop in an accumulation layer.

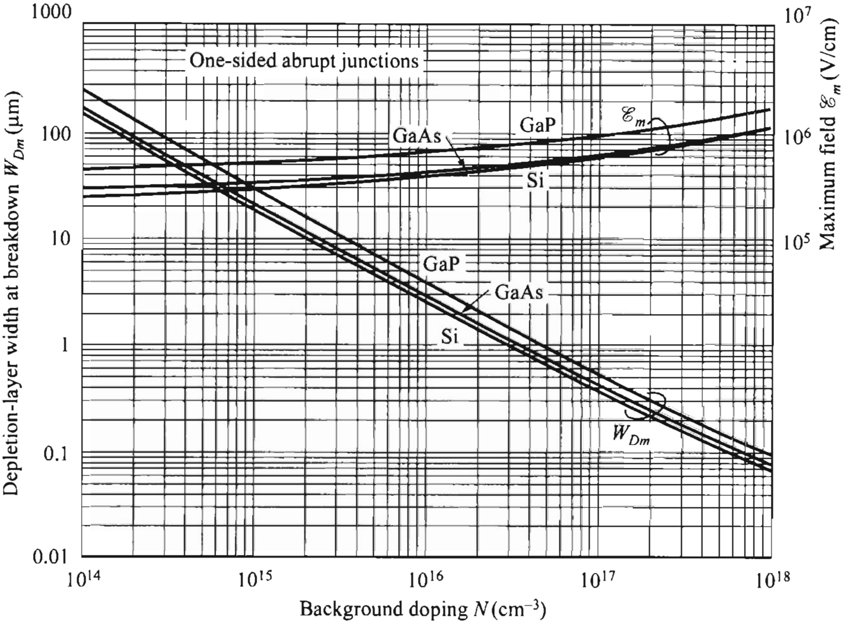


(a)

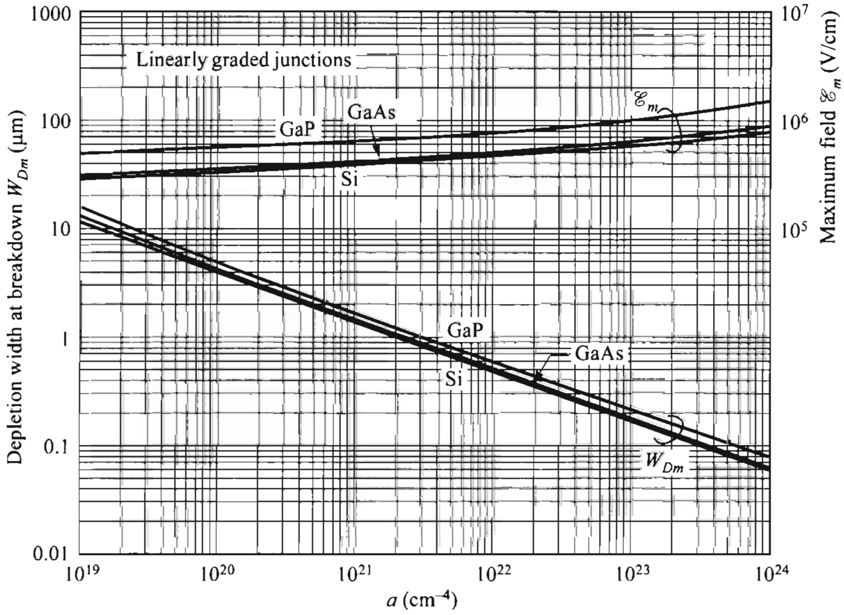


(b)

Fig. 16 Avalanche breakdown voltage in Si, ⟨100⟩-oriented GaAs, and GaP, for (a) one-sided abrupt junctions (vs. impurity concentration) and (b) linearly graded junctions (vs. impurity gradient). The dashed lines indicate the maximum doping or doping gradient beyond which tunneling will dominate the breakdown characteristics. (After Ref. 14.)



(a)



(b)

Fig. 17 Depletion-layer width and maximum field at breakdown in Si, ⟨100⟩-oriented GaAs, and GaP for (a) one-sided abrupt junctions and (b) linearly graded junctions. (After Ref. 14.)

An approximate universal expression can be given as follows for the results above comprising all semiconductors studied:

$$V_{BD} \approx 60\left(\frac{E_g}{1.1\ \text{eV}}\right)^{3/2}\left(\frac{N}{10^{16}\ \text{cm}^{-3}}\right)^{-3/4} \quad \text{V} \tag{104}$$

for abrupt junctions where E_g is the room-temperature bandgap in eV, and N is the background doping in cm^{-3}; and

$$V_{BD} \approx 60\left(\frac{E_g}{1.1\ \text{eV}}\right)^{6/5}\left(\frac{a}{3\times10^{20}\ \text{cm}^{-4}}\right)^{-2/5} \quad \text{V} \tag{105}$$

for linearly graded junctions where a is the impurity gradient in cm^{-4}.

For diffused junctions with a linear gradient near the junction and a constant doping on one side (Fig. 18 inset), the breakdown voltage lies between the two limiting cases considered previously[18] (Fig. 16). As shown in Fig. 18, for large a, the breakdown voltage of these junctions is given by the abrupt junction results (bottom line); on the other hand, for small a, V_{BD} will be given by the linearly graded junction results (parallel lines) and is independent of N_B.

In Figs. 16 and 17, it is assumed that the semiconductor layer is thick enough to support the maximum depletion-layer width W_{Dm} at breakdown. If, however, the semiconductor layer W is smaller than W_{Dm} (shown in Fig. 19, inset), the device will be punched through (i.e., the depletion layer reaches the n^+ substrate) prior to breakdown. As the reverse bias increases further, the depletion width cannot continue to expand and the device will break down prematurely. The maximum electric field $\mathscr{E}_m$

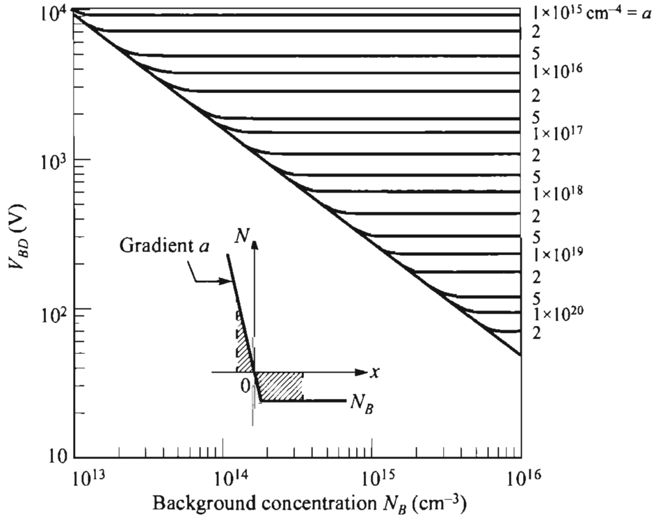


Fig. 18 Breakdown voltage for Si diffused junctions at 300 K. The inset shows the space-charge distribution. (After Ref. 18.)

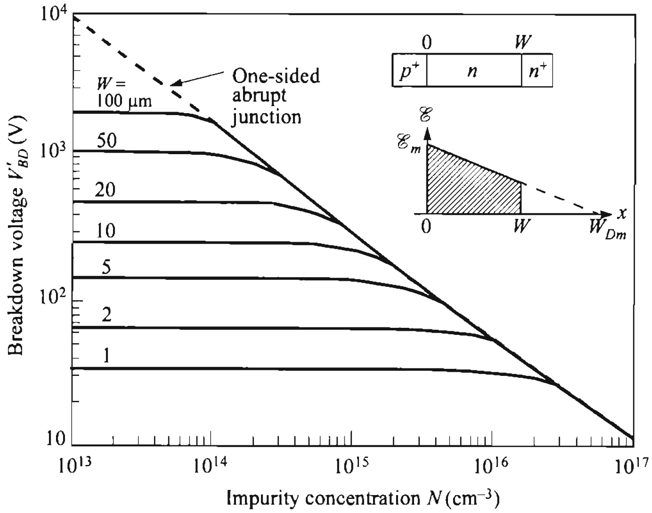


Fig. 19 Breakdown voltage for Si p^+-π-n^+ and p^+-ν-n^+ junctions, where π stands for lightly doped p-type and ν for lightly doped n-type. W is the thickness of the π- or ν-region.

is essentially the same as for the nonpunched-through diode. Therefore, the reduced breakdown voltage V'_{BD} for the punched-through diode, compared to a regular device with V_{BD} for the same doping, can be given by

$$\frac{V'_{BD}}{V_{BD}} = \frac{\text{Shaded area in figure insert}}{(\mathscr{E}_m W_{Dm})/2} = \left(\frac{W}{W_{Dm}}\right)\left(2 - \frac{W}{W_{Dm}}\right) . \tag{106}$$

Punch-through usually occurs when the doping concentration N becomes sufficiently low as in a p^+-π-n^+ or p^+-ν-n^+ diode, where π stands for a lightly doped p-type and ν for a lightly doped n-type semiconductor. The breakdown voltages for such diodes as calculated from Eq. 106 are shown in Fig. 19 as a function of the background doping for Si one-sided abrupt junction formed on epitaxial substrates (e.g., ν on n^+ with the epitaxial-layer thickness W as a parameter). For a given thickness, the breakdown voltage approaches a constant value as the doping decreases, corresponding to the punch-through of the epitaxial layer.

The results shown so far are for avalanche breakdowns at room temperature. At higher temperatures the breakdown voltage increases. A qualitative explanation of this increase is that hot carriers passing through the depletion layer under a high field lose part of their energy to optical phonons via scattering, resulting in a smaller ionization rate (see Fig. 24 of Chapter 1). Therefore, the carriers lose more energy to the crystal lattice along a given distance at a constant field. Hence, the carriers must pass through a greater potential difference (or higher voltage) before they can acquire sufficient energy to generate an electron-hole pair. The predicted values of V_{BD} normalized to the room-temperature value are shown in Fig. 20 for silicon. Note that there are substantial increases of the breakdown voltage, especially for lower dopings (or small gradient) at higher temperatures.[20]

Edge Effects. For junctions formed by a planar process, a very important junction curvature effect at the perimeter should be considered. A schematic diagram of a planar junction is shown in Fig. 21a. Note that at the perimeter, the depletion region is narrower and the field is higher. Since the cylindrical and/or spherical regions of the junction have a higher field intensity, the avalanche breakdown voltage is determined by these regions. The potential $\psi(r)$ and the electric field $\mathscr{E}(r)$ in a cylindrical or spherical *p-n* junction can be calculated from Poisson equation:

$$\frac{1}{r^n}\frac{d}{dr}[r^n\mathscr{E}(r)] = \frac{\rho(r)}{\varepsilon_s} \tag{107}$$

where n equals 1 for the cylindrical junction, and 2 for the spherical junction. The solution for $\mathscr{E}(r)$ can be obtained from this equation and is given by

$$\mathscr{E}(r) = \frac{1}{\varepsilon_s r^n}\int_{r_j}^{r} r^n\rho(r)dr + \frac{C_1}{r^n} \tag{108}$$

where r_j is the radius of curvature of the metallurgical junction, and the constant C_1 must be adjusted so that the integration of the field is equal to the built-in potential.

The calculated results for Si one-sided abrupt junctions at 300 K can be expressed by a simple equation:[18]

$$\frac{V_{CY}}{V_{BD}} = \left[\frac{1}{2}(\eta^2 + 2\eta^{6/7})\ln(1 + 2\eta^{-8/7}) - \eta^{6/7}\right] \tag{109}$$

for cylindrical junctions, and

$$\frac{V_{SP}}{V_{BD}} = [\eta^2 + 2.14\eta^{6/7} - (\eta^3 + 3\eta^{13/7})^{2/3}] \tag{110}$$

for spherical junctions, where V_{CY} and V_{SP} are the breakdown voltages of cylindrical and spherical junctions, respectively, V_{BD} and W_{Dm} are the breakdown voltage and maximum depletion width of a plane junction having the same background doping, and $\eta \equiv r_j/W_{Dm}$. Figure 21b illustrates the numerical results as a function of η. Clearly, as the radius of curvature becomes smaller, so does the breakdown voltage. However, for linearly graded cylindrical or spherical junctions, the calculated results show that the breakdown voltage is relatively independent of its radius of curvature.[21]

Another edge effect that causes premature breakdown is due to an MOS (metal-oxide semiconductor) structure over the junction at the surface. Such a configuration is often called a gated diode. At certain gate biases, the field near the gate edge is higher than in the planar portion of the junction and breakdown changes location from the surface area of the metallurgical junction to the edge of the gate. This gate-voltage dependence of breakdown is shown in Fig. 22. At high positive gate bias on a p^+-n junction, the p^+-surface is depleted while the n-surface is accumulated. Breakdown occurs near the metallurgical junction at the surface. As the gate bias is swept

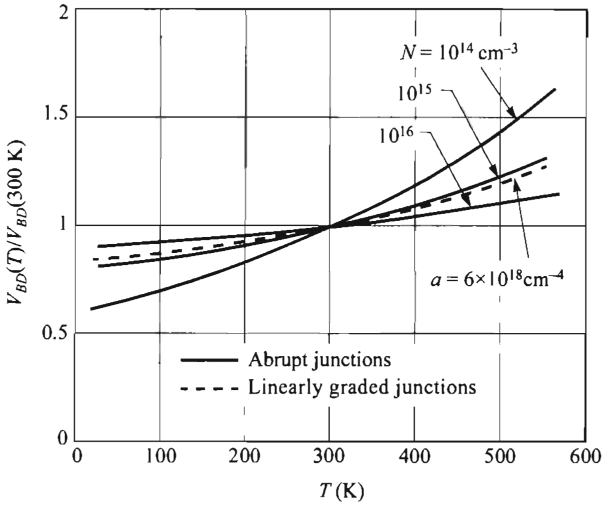


Fig. 20 Normalized avalanche breakdown voltage versus lattice temperature, in silicon. The breakdown voltage generally increases with temperature. (After Ref. 19.}

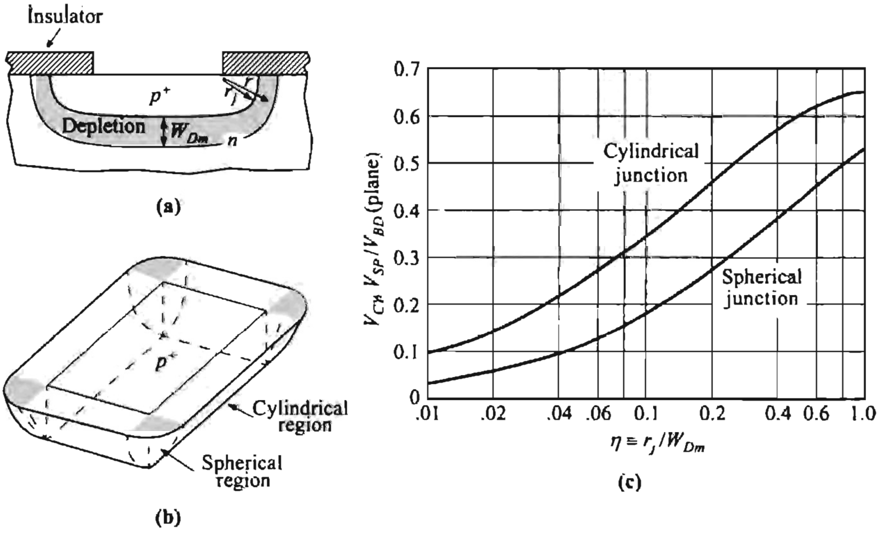


Fig. 21 (a) A planar diffusion or implantation process forms a junction curvature near the edges of the mask with r_j the radius of curvature. (b) Three-dimensional view of the junction curvature showing the spherical region at the corners. (c) Normalized breakdown voltage of cylindrical and spherical junctions as a function of the normalized radius of curvature. (After Ref. 18.)

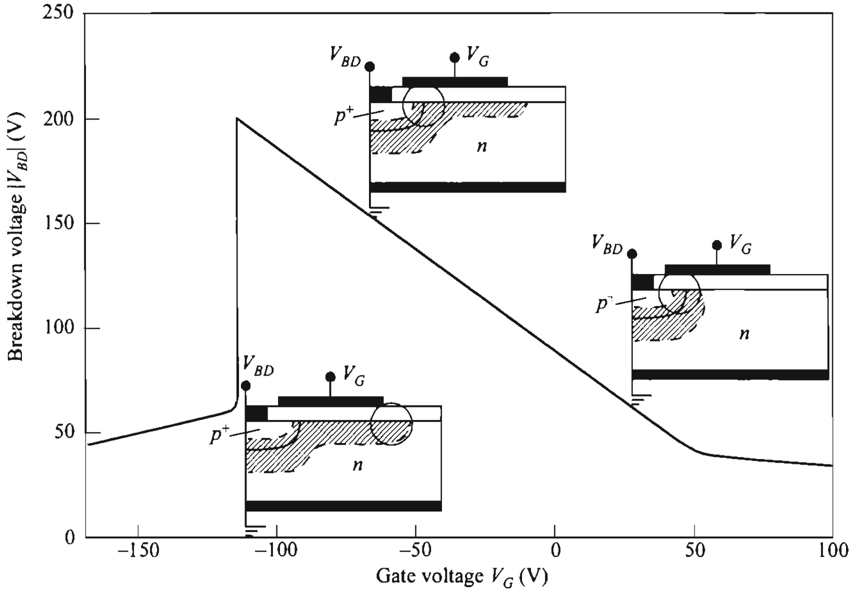


Fig. 22 Gate-voltage dependence of breakdown in a gated diode. The location of high-field breakdown shifts with gate bias. (After Ref. 22.)

more negatively, the location of breakdown moves toward the *n*-side (to the right). In the middle gate-bias range, the breakdown voltage has a linear dependence on the gate bias, with[23]

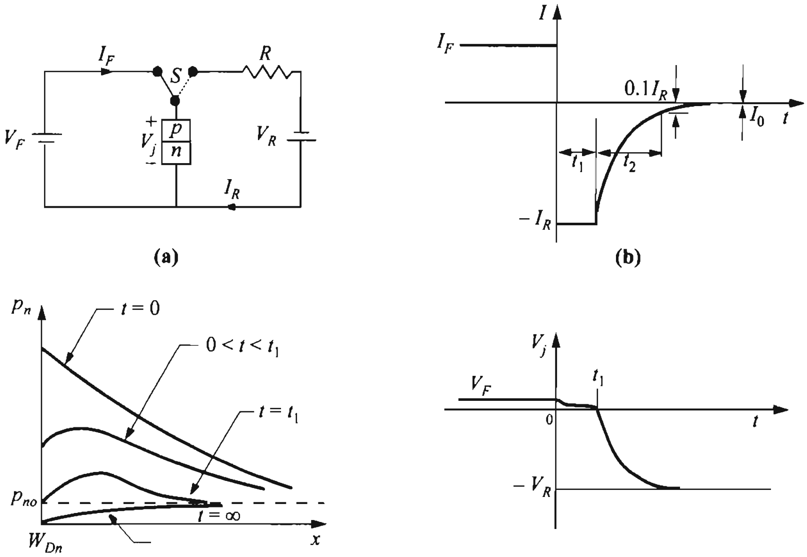
$$V_{BD} = mV_G + \text{constant} \tag{111}$$

and $m \leq 1$. At some high negative gate bias, the field directly under the gate edge is high enough to cause breakdown, and the breakdown voltage collapses. This gated-diode breakdown phenomenon is reversible and the measurement can be repeated. To minimize this edge effect, the oxide thickness should be above a critical value.[22] This mechanism is also responsible for the gate-induced drain leakage (GIDL) of the MOSFET (see Section 6.4.5).

2.5 TRANSIENT BEHAVIOR AND NOISE

2.5.1 Transient Behavior

For switching applications the transitions from forward bias to reverse bias and vice versa must be nearly abrupt and the transient time short. For a *p*-*n* junction, while the latter is reasonably fast, the response from forward to reverse is limited by minority-carrier charge storage. Figure 23a shows a simple circuit in which a forward current I_F flows in the *p*-*n* junction; at time $t = 0$, the switch S is suddenly thrown to the right,



(a) (b)

(c) (d)

Fig. 23 Transient behavior of a *p-n* junction. (a) Basic switching circuit. (b) Transient current response. (c) Minority-carrier distribution outside depletion edge for various time intervals. (d) Transient junction-voltage response. (After Ref. 24.)

and an initial reverse current $I_R = (V_R - V_F)/R$ flows. The transient time is defined as the time in which the current drops to 10% of the initial reverse current I_R, and is equal to the sum of t_1 and t_2 as shown in Fig. 23b, where t_1 and t_2 are the time intervals for the constant-current phase and the decay phase, respectively.

Consider the constant-current phase (also called storage phase) first. The continuity equation as given in Chapter 1 can be written for the *n*-type side of a p^+-n junction ($p_{po} \gg n_{no}$) as

$$\frac{\partial p_n(x, t)}{\partial t} = D_p \frac{\partial^2 p_n(x, t)}{\partial x^2} - \frac{p_n(x, t) - p_{no}}{\tau_p}. \tag{112}$$

The boundary conditions are that at $t = 0$ the initial distribution of holes is a steady-state solution to the diffusion equation, and that under forward bias the voltage across the junction is given from Eq. 53b as

$$V_j(t) = \frac{kT}{q} \ln\left[\frac{p_n(0, t)}{p_{no}}\right]. \tag{113}$$

The distribution of the minority-carrier density p_n with time is shown in Fig. 23c. From Eq. 113 it can be calculated that, as long as $p_n(0,t)$ is greater than p_{no} (in the interval $0 < t < t_1$), the junction voltage V_j remains of the order of kT/q, as shown in Fig. 23d. In this time interval the reverse current is approximately constant and we have the constant-current phase. The solution of the time-dependent continuity equation gives t_1 by the transcendental equations[24]

$$\operatorname{erf}\sqrt{\frac{t_1}{\tau_p}} = \frac{1}{1+(I_R/I_F)}. \tag{114}$$

However, an explicit expression for t_1 can be obtained, using a charge-control model which can also provide some insight into the problem. The stored minority-carrier charge in the lightly doped side is given by the integral

$$Q_s = qA\int \Delta p_n dx. \tag{115}$$

Integration of the continuity equation, after the current is switched to the reversed mode, becomes

$$-I_R = \frac{dQ_s}{dt} + \frac{Q_s}{\tau_p}. \tag{116}$$

With the initial condition given by the forward current $Q_s(0) = I_F\tau_p$, the solution is given by

$$Q_s(t) = \tau_p\left[-I_R + (I_F + I_R)\exp\left(\frac{-t}{\tau_p}\right)\right]. \tag{117}$$

By setting $Q_s = 0$, t_1 can be obtained as

$$t_1 = \tau_p \ln\left(1 + \frac{I_F}{I_R}\right). \tag{118}$$

A comparison of Eq. 118 to the exact solution of Eq. 114 shows that this estimate gives higher values by a factor of ≈ 2 for $I_F/I_R = 0.1$ and ≈ 20 for $I_F/I_R = 10$.

After t_1, the hole density starts to decrease below its equilibrium value p_{no}. The junction voltage tends to reach $-V_R$ and a new boundary condition now holds. This phase is the decay phase with the initial boundary condition $p_n(0,t_1) = p_{no}$. The solution for t_2 is given by another transcendental equation

$$\operatorname{erf}\sqrt{\frac{t_2}{\tau_p}} + \frac{\exp(-t_2/\tau_p)}{\sqrt{\pi t_2/\tau_p}} = 1 + 0.1\left(\frac{I_R}{I_F}\right). \tag{119}$$

The total results for t_1 and t_2 are shown in Fig. 24 where the solid lines are for the plane junction with the length of the n-type material W much greater than the diffusion length ($W \gg L_p$), and the dashed lines are for the narrow-base junction with $W \ll L_p$. For a large I_R/I_F ratio, the transient time can be approximated by

$$t_1 + t_2 \approx \frac{\tau_p}{2}\left(\frac{I_R}{I_F}\right)^{-2} \tag{120}$$

for $W \gg L_p$, or

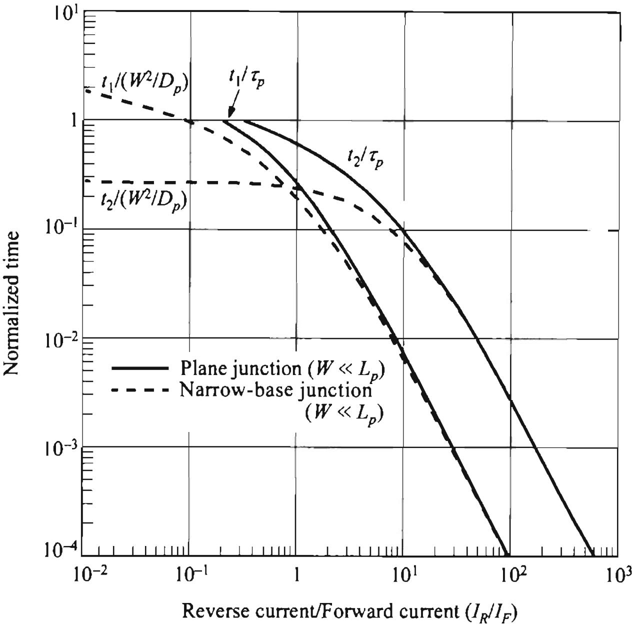


Fig. 24 Normalized time versus the ratio of reverse current to forward current. W is width of the n-region in a p^+-n junction. (After Ref. 24.)

$$t_1 + t_2 \approx \frac{W^2}{2D_p}\left(\frac{I_R}{I_F}\right)^{-2} \tag{121}$$

for $W \ll L_p$. For example, if one switches a junction (of $W \gg L_p$) from forward 10 mA to reverse 10 mA ($I_R/I_F = 1$), the time for the constant-current phase is $0.3\tau_p$, and that for the decay phase is about $0.6\tau_p$. Total transient time is then $0.9\tau_p$. A fast switch requires that τ_p be small for all cases. The lifetime τ_p can be substantially reduced by introducing impurities with deep levels in the forbidden gap, such as gold in silicon.

2.5.2 Noise

The term "noise" refers to spontaneous fluctuations in the current passing through, or the voltage developed across, semiconductor bulk materials or devices. Since semiconductor devices are mainly used to amplify small signals or to measure small physical quantities, spontaneous fluctuations in current or voltage set a lower limit to these signals. It is important to know the factors contributing to these limits, to use this knowledge to optimize operating conditions, and to find new methods and new technologies to reduce noise.

Observed noise is generally classified into (1) thermal noise or Johnson noise, (2) flicker noise, and (3) shot noise. Thermal noise occurs in any conductor or semiconductor device and is caused by the random thermal motion of the current carriers. It is also called white noise because its level is the same at all frequencies. The open-circuit mean-square voltage of thermal noise is given by[25,26]

$$\langle V_n^2 \rangle = 4kTBR \tag{122}$$

where B is the bandwidth in Hz, and R the real part of the dynamic impedance (dV/dI) between terminals. At room temperature, for a semiconductor device with 1 kΩ resistance, the root-mean-square voltage $\sqrt{\langle V_n^2 \rangle}$ measured with a 1-Hz bandwidth is only about 4 nV.

Flicker noise is distinguished by its peculiar spectral distribution which is proportional to $1/f^{\alpha}$ with α generally close to unity (the so-called $1/f$ noise). Flicker noise is thus important at lower frequencies. For most semiconductor devices, the origin of flicker noise is the surface effect. The $1/f$ noise-power spectrum has been correlated both qualitatively and quantitatively with the lossy part of the metal-insulator-semiconductor (MIS) gate impedance due to carrier recombination at the interface traps.

Shot noise is due to the discreteness of charge carriers that contribute to current flow, and it constitutes the major noise in most semiconductor devices. It is independent of frequency (white spectrum) at low and intermediate frequencies. At higher frequencies the shot-noise spectrum also becomes frequency-dependent. The mean-square noise current of shot noise for a *p-n* junction is given by

$$\langle i_n^2 \rangle = 2qB|I| \tag{123}$$

where I can be forward or reverse current. For low injection the total mean-square noise current (neglecting $1/f$ noise) is the sum

$$\langle i_n^2 \rangle = \frac{4kTB}{R} + 2qB|I| \,. \tag{124}$$

From the Shockley equation we obtain

$$\frac{1}{R} = \frac{dI}{dV} = \frac{d}{dV}\left\{ I_0\left[\exp\left(\frac{qV}{kT}\right) - 1\right]\right\} = \frac{qI_0}{kT}\exp\left(\frac{qV}{kT}\right) . \tag{125}$$

Substituting Eq. 125 into Eq. 124 yields for the forward-bias condition,

$$\begin{aligned} \langle i_n^2 \rangle &= 4qI_0B\exp\left(\frac{qV_F}{kT}\right) + 2qI_0B\left[\exp\left(\frac{qV_F}{kT}\right) - 1\right] \\ &\approx 6qI_0B\exp\left(\frac{qV_F}{kT}\right) \,. \end{aligned} \tag{126}$$

Experimental measurements indeed confirm that the mean-square noise current is proportional to the saturation current I_0, which can be increased by irradiation.

2.6 TERMINAL FUNCTIONS

A *p-n* junction is a two-terminal device that can perform various terminal functions, depending upon its biasing condition as well as its doping profile and device geometry. In this section we discuss briefly some interesting device performances based on its current-voltage, capacitance-voltage, and breakdown characteristics discussed in previous sections. Many other related two-terminal devices will be considered in subsequent chapters (e.g., tunnel diode in Chapter 8 and IMPATT diode in Chapter 9).

2.6.1 Rectifier

A rectifier is a two-terminal device that gives a very low resistance to current flow in one direction and a very high resistance in the other direction, i.e., it allows current in only one direction. The forward and reverse resistances of a rectifier can be derived from the current-voltage relationship of a practical diode,

$$I = I_0\left[\exp\left(\frac{qV}{\eta kT}\right) - 1\right] \tag{127}$$

where I_0 is the saturation current and the ideality factor η generally has a value between 1 (for diffusion current) and 2 (for recombination current). The forward dc (or static) resistance R_F and small-signal (or dynamic) resistance r_F are obtainable from Eq. 127:

$$R_F \equiv \frac{V_F}{I_F} \approx \frac{V_F}{I_0}\exp\left(\frac{-qV_F}{\eta kT}\right), \tag{128}$$

$$r_F \equiv \frac{dV_F}{dI_F} \approx \frac{\eta kT}{qI_F}. \tag{129}$$

The reverse dc resistance R_R and small-signal resistance r_R are given by

$$R_R \equiv \frac{V_R}{I_R} \approx \frac{V_R}{I_0}, \tag{130}$$

$$r_R \equiv \frac{dV_R}{dI_R} = \frac{\eta kT}{qI_0}\exp\left(\frac{q|V_R|}{\eta kT}\right). \tag{131}$$

Comparing Eqs. 128–131 shows that the dc rectification ratio R_R/R_F varies with the factor $(V_R/V_F)\exp(qV_F/\eta kT)$, while the ac rectification ratio r_R/r_F varies with $(I_F/I_0)\exp(q|V_R|/\eta kT)$.

p-n junction rectifiers generally have slow switching speeds; that is, a significant time delay is necessary to obtain high impedance after switching from the forward-conduction state to the reverse-blocking state. This time delay (proportional to the minority-carrier lifetime as shown in Fig. 24) is of little consequence in rectifying 60-Hz currents. For high-frequency applications, the lifetime should be sufficiently reduced to maintain rectification efficiency. The majority of rectifiers have power-dissipation capabilities from 0.1 to 10 W, reverse breakdown voltages from 50 to 2500 V (for a high-voltage rectifier two or more *p-n* junctions are connected in series), and switching times from 50 ns for low-power diodes to about 500 ns for high-power diodes.

A rectifier has many circuit applications.[27] It is used to transform ac signals into different specials waveforms. Examples are half-wave and full-wave rectifiers, clipper and clamper circuits, peak detector (demodulator), etc. It can also be used as a ESD (electrostatic discharge) protection device.

2.6.2 Zener Diode

A Zener diode (also called voltage regulator) has a well-controlled breakdown voltage, called the Zener voltage, with sharp breakdown characteristics in the reverse-bias region. Prior to breakdown, the diode has a very high resistance; after breakdown the diode has a very small dynamic resistance. The terminal voltage is thus limited (or regulated) by the breakdown voltage, and this is used to establish a fixed reference voltage.

Most Zener diodes are made of Si, because of the low saturation current in Si diodes and the advanced Si technology. They are special *p-n* junctions with higher doping concentrations on both sides. As discussed in Section 2.4, for breakdown voltage V_{BD} larger than $6E_g/q$ (≈ 7 V for Si), the breakdown mechanism is mainly avalanche multiplication, and the temperature coefficient of V_{BD} is positive. For $V_{BD} < 4E_g/q$ (≈ 5 V for Si), the breakdown mechanism is band-to-band tunneling, and the temperature coefficient of V_{BD} is negative. For $4E_g/q < V_{BD} < 6E_g/q$, the breakdown is due to a combination of these two mechanisms. One can connect, for example, a negative-temperature-coefficient diode in series with a positive-temperature-coefficient diode to produce a temperature-independent regulator (with a temperature coefficient of the order of 0.002% per °C), which is suitable as a voltage reference.

2.6.3 Varistor

A varistor (variable resistor) is a two-terminal device that shows nonohmic behavior, i.e., voltage-dependent resistance.[28] Equations 128 and 129 show the nonohmic characteristics of a *p-n* junction diode in the forward-bias region. Similar nonohmic characteristics are obtainable from metal-semiconductor contacts considered in Chapter 3. An interesting application of varistors is their use as a symmetrical fractional-voltage (≈ 0.5 V) limiter by connecting two diodes in parallel, oppositely poled. The two-diode unit will exhibit the forward *I-V* characteristics in either direction. A varistor, being a nonlinear device, is also useful in microwave modulation, mixing, and detection (demodulation). Varistors based on metal-semiconductor contacts are more common due to their higher speed from the absence of minority-charge storage.

2.6.4 Varactor

The term *varactor* comes from *variable reactor* and means a device whose reactance (or capacitance) can be varied in a controlled manner with a dc bias voltage. Varactor diodes are widely used in parametric amplification, harmonic generation, mixing, detection, and voltage-variable tuning.

For this application, the forward bias is to be avoided because of excessive current which is undesirable for any capacitor. The basic capacitance-voltage relationships in reverse bias have already been derived in Section 2.2. We shall now extend the previous derivations of abrupt and linearly graded doping distributions to a more general case. The one-dimensional Poisson equation is given as

$$\frac{d^2\psi_i}{dx^2} = -\frac{qN}{\varepsilon_s} \tag{132}$$

where N is the generalized doping distribution (negative sign for donors) as shown in Fig. 25a (assuming one side is heavily doped):

$$N = Bx^m \quad \text{for} \quad x \geq 0. \tag{133}$$

For $m = 0$ we have $N = B$, corresponding to the uniformly doped (or one-sided abrupt junction) case. For $m = 1$, the doping profile corresponds to a one-sided linearly graded case. For $m < 0$, the device is called a "hyper-abrupt" junction. The hyperabrupt doping profile can be achieved by an epitaxial process or by ion implantation. The boundary conditions are $\psi_i(x = 0) = 0$ and $\psi_i(x = W_D) = V_R + \psi_{bi}$, where V_R is the applied reverse voltage and ψ_{bi} is the built-in potential. Integrating the Poisson equation with the boundary conditions, we obtain for the depletion-layer width and the differential capacitance per unit area[29]

$$W_D = \left[\frac{\varepsilon_s(m+2)(V_R + \psi_{bi})}{qB}\right]^{1/(m+2)}, \tag{134}$$

$$C_D \equiv \frac{\varepsilon_s}{W_D} = \left[\frac{qB\varepsilon_s^{m+1}}{(m+2)(V_R + \psi_{bi})}\right]^{1/(m+2)} \propto (V_R + \psi_{bi})^{-s}, \tag{135}$$

$$s \equiv \frac{1}{m+2}. \tag{136}$$

One important parameter in characterizing the varactor is the sensitivity defined by[30]

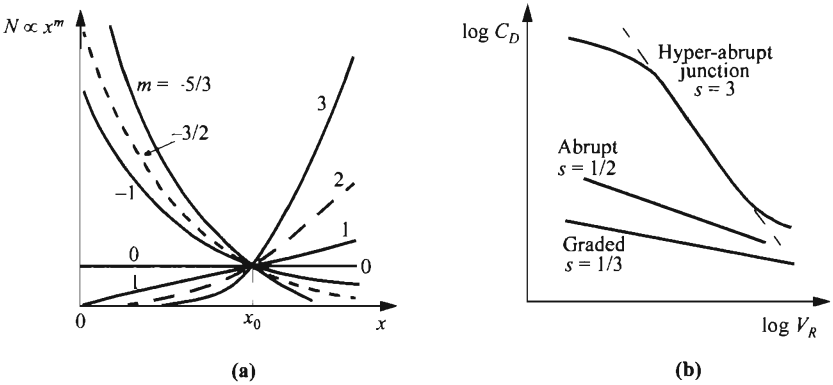


Fig. 25 (a) Various impurity distributions (normalized at x_0) for varactors. (b) Log-log plot of depletion-layer capacitance versus reverse bias. (After Refs. 29 and 30.)

$$-\frac{dC_D}{C_D}\frac{V_R}{dV_R} = -\frac{d(\log C_D)}{d(\log V_R)} = \frac{1}{m+2} = s. \qquad (137)$$

The larger the s, the larger will be the capacitance variation with biasing voltage. For linearly graded junctions, $m = 1$ and $s = 1/3$; for abrupt junctions, $m = 0$ and $s = 1/2$; for hyper-abrupt junction with $m = -1, -3/2$, or $-5/3$, the value of s is 1,2, or 3, respectively. The capacitance-voltage relationships for these junction diodes are shown in Fig. 25b. The hyper-abrupt junction, as expected, has the highest sensitivity and gives rise to the largest capacitance variation.

2.6.5 Fast-Recovery Diode

Fast-recovery diodes are designed to give ultrahigh switching speed. The devices can be classified into two types: *p-n* junction diodes and metal-semiconductor diodes. The general switching behavior of both types can be described by Fig. 23b. The total recovery time ($t_1 + t_2$) for a *p-n* junction diode can be substantially reduced by introducing recombination centers, such as Au in Si, to reduce the carrier lifetime. Although the recovery time is directly proportional to the lifetime τ, as shown in Fig. 24, it is not possible, unfortunately, to reduce the recovery time indefinitely by introducing an extremely large number of recombination centers N_t, because the reverse generation current of a *p-n* junction is proportional to N_t (Eqs. 66 and 67). For direct bandgap semiconductors, such as GaAs, the minority-carrier lifetimes are generally much smaller than that of Si. This results in ultra-high-speed GaAs *p-n* junction diodes with recovery times of the order of 0.1 ns or less. For Si the practical recovery time is in the range of 1 to 5 ns.

The metal-semiconductor diodes (Schottky diodes) fundamentally exhibit ultrahigh-speed characteristics, because they are majority-carrier devices and the minority-carrier storage effect is negligible. We discuss metal-semiconductor contacts in detail in Chapter 3.

2.6.6 Charge-Storage Diode

In contrast to fast-recovery diodes, a charge-storage diode is designed to store a charge while conducting in the forward direction and, upon switching to the reverse direction, to conduct a reverse current for a short period. A particularly interesting charge-storage diode is the step-recovery diode (also called the snapback diode) that conducts in the reverse direction for a short period and then abruptly cuts off the current as the stored charge has been dissipated. In other words, it is desirable here to reduce the decay phase or t_2 without shortening the storage phase or t_1. Most charge-storage diodes are made from Si with relatively long minority-carrier lifetimes ranging from 0.5 to 5 μs. Note that the lifetimes are about 1000 times longer than for fast-recovery diodes. The mechanism to reduce the decay phase is by a special doping profile such that the injected charge is confined closer to the junction. This cutoff occurs in the range of picoseconds and results in a fast-rising wavefront which is rich in harmonics. Because of these characteristics, step-recovery diodes are used in harmonic generation and pulse shaping.

2.6.7 *p-i-n* Diode

A *p-i-n* diode is a *p-n* junction with an intrinsic layer (*i*-region) sandwiched between the *p*-layer and the *n*-layer. In practice, however, the idealized *i*-region is approximated by either a high-resistivity *p*-layer (referred to as π-layer) or a high-resistivity *n*-layer (ν-layer). The *p-i-n* diode has found wide applications in microwave circuits. Its special feature is a wide intrinsic layer that provides unique properties such as low and constant capacitance, high breakdown voltage in reverse bias, and most interestingly, as a variolosser (variable attenuator) by controlling the device resistance which varies approximately linearly with the forward bias current. The switching time is approximately given by $W/2\upsilon_s$, where W is the width of the *i*-region.[31] It can modulate signals up to the GHz range. Furthermore, the forward characteristics of a thyristor (refer to Chapter 11) in its on-state closely resemble those of a *p-i-n* diode.

At near zero or low reverse bias, the lightly doped intrinsic layer starts to be fully depleted, and the capacitance is given by

$$C = \frac{\varepsilon_s}{W}. \tag{138}$$

Once fully depleted, its capacitance is independent of reverse bias. Figure 19 gives the breakdown voltage of a *p-i-n* diode under reverse bias. Since there is little net charge within the intrinsic layer, the electric field is constant and the breakdown voltage can be estimated by

$$V_{BD} \approx \mathscr{E}_m W \tag{139}$$

where the maximum breakdown field $\mathscr{E}_m$ for Si at lower dopings is about 2.5×10^5 V/cm. These two equations show that the width of the *i*-region W controls the trade-off between frequency response and power (from maximum voltage).

Under forward conditions, holes are injected from the *p*-region and electrons from the *n*-region. As the injected carrier densities are nearly equal (and uniform) due to charge neutrality, they are much higher than the *i*-region doping concentration, so the *p-i-n* diode is generally operated in the high-injection condition, $\Delta p = \Delta n \gg n_i$. The current conduction is via recombination within the *i*-region and is given by (see Eq. 74)

$$J_{re} = \int_0^W qUdx = \frac{qWn_i}{2\tau}\exp\left(\frac{qV_F}{2kT}\right). \tag{140}$$

For a detailed discussion of the dc *I-V* characteristics, the readers are referred to Section 11.2.4.

The most interesting phenomenon for a *p-i-n* diode, however, is for small signals at high frequencies ($> 1/2\pi\tau$) at which the stored carriers within the intrinsic layer are not completely swept away by the RF signal or by recombination. At these frequencies there is no rectification and the *p-i-n* diode behaves like a pure resistor whose value is determined solely by the injected charge, proportional to the dc bias current. This dynamic RF resistance is simply given by

$$R_{RF} = \rho\frac{W}{A} = \frac{W}{q\Delta n(\mu_n + \mu_p)A}$$
$$= \frac{W^2}{J_F \tau(\mu_n + \mu_p)A} \; . \tag{141}$$

Here the relationship $J_F = qW\Delta n/\tau$ has been assumed. The RF resistance is controlled by the dc bias current, and typical characteristics are shown in Fig. 26.

2.7 HETEROJUNCTIONS

Some properties of heterojunctions have been discussed in Section 1.7. When the two semiconductors have the same type of conductivity, the junction is called an *isotype* heterojunction. When the conductivity types differ, the junction is called an *anisotype* heterojunction which is a much more useful and common structure than its counterpart. In 1951, Shockley proposed the abrupt heterojunction to be used as an efficient emitter-base injector in a bipolar transistor.[33] In the same year, Gubanov published a theoretical paper on heterojunctions.[34] Kroemer later analyzed a similar, although graded, heterojunction as a wide-bandgap emitter.[35] Since then, heterojunctions have been extensively studied, and many important applications have been made, among them the room-temperature injection laser, light-emitting diode (LED), photodetector, and solar cell, to name a few. In many of these applications, by forming periodic heterojunctions with layer thickness of the order of 10 nm, we utilize the interesting properties of quantum wells and superlattices. Additional information on heterojunctions can be found in Refs. 36–39.

2.7.1 Anisotype Heterojunction

The energy-band model of an idealized anisotype abrupt heterojunction without interface traps was proposed by Anderson[40] based on the previous work of Shockley. We consider this model next, since it can adequately explain most transport processes,

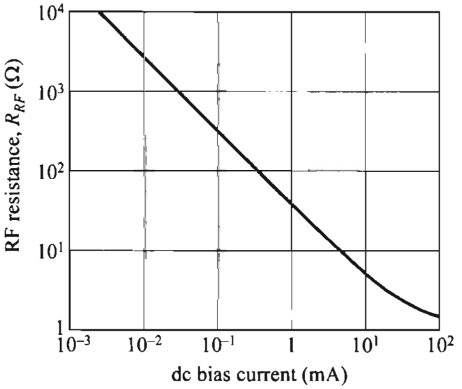


Fig. 26 Typical RF resistance as a function of dc forward current. (After Ref. 32.)

and only slight modification of the model is needed to account for nonideal cases such as interface traps. Figures 27a and c show the energy-band diagrams of two isolated semiconductors of opposite types. The two semiconductors are assumed to have different bandgaps E_g, different permittivities ε_s, different work functions ϕ_m, and different electron affinities χ. Work function and electron affinity are defined as the energy required to remove an electron from the Fermi level E_F and from the bottom of the conduction band E_C, respectively, to a position just outside the material (vacuum level). The difference in energy of the conduction-band edges in the two semiconductors is represented by ΔE_C and that in the valence-band edges by ΔE_V. The

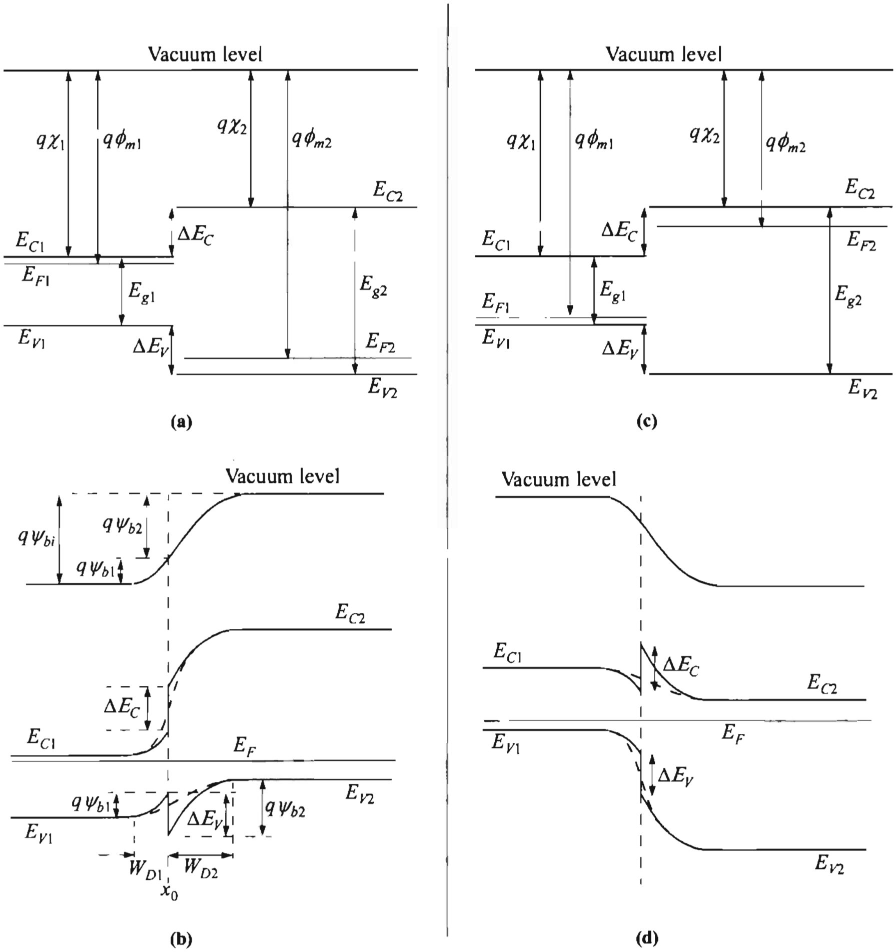

Vacuum level
$q\chi_1$
$q\phi_{m1}$
$q\chi_2$
$q\phi_{m2}$
E_{C2}
E_{C1}
ΔE_C
E_{F1}
E_{g1}
E_{g2}
E_{V1}
ΔE_V
E_{F2}
E_{V2}
(a)
Vacuum level
$q\chi_1$
$q\phi_{m1}$
$q\chi_2$
$q\phi_{m2}$
E_{C2}
E_{C1}
ΔE_C
E_{F2}
E_{g1}
E_{g2}
E_{F1}
E_{V1}
ΔE_V
E_{V2}
(c)
Vacuum level
$q\psi_{bi}$
$q\psi_{b2}$
$q\psi_{b1}$
E_{C2}
ΔE_C
E_{C1}
E_F
$q\psi_{b1}$
ΔE_V
$q\psi_{b2}$
E_{V2}
E_{V1}
W_{D1}
W_{D2}
x_0
(b)
Vacuum level
E_{C1}
ΔE_C
E_{C2}
E_F
E_{V1}
ΔE_V
E_{V2}
(d)

Fig. 27 Energy-band diagrams for (a) two isolated semiconductors of opposite types and different E_g (of which the smaller bandgap is *n*-type) and (b) their idealized anisotype heterojunction at thermal equilibrium. In (c) and (d), the smaller bandgap is *p*-type. In (b) and (d), the dashed lines across the junctions represent graded composition. (After Ref. 40.)

electron affinity rule ($\Delta E_C = q\Delta\chi$) shown in Figure 27 may not be a valid assumption in all cases. However, by choosing ΔE_C as an empirical quantity, the Anderson model remains satisfactory and unaltered.[41]

When a junction is formed between these semiconductors, the energy-band profile at equilibrium is as shown in Fig. 27b for an *n-p* anisotype heterojunction where, in this example, the narrow-bandgap material is *n*-type. Since the Fermi level must coincide on both sides in equilibrium and the vacuum level is everywhere parallel to the band edges and is continuous, the discontinuity in the conduction-band edges (ΔE_C) and valence-band edges (ΔE_V) is invariant with doping in those cases where E_g and χ are not functions of doping (i.e., nondegenerate semiconductors). The total built-in potential ψ_{bi} is equal to the sum of the partial built-in voltages ($\psi_{b1} + \psi_{b2}$), where ψ_{b1} and ψ_{b2} are the electrostatic potentials supported at equilibrium by semiconductors 1 and 2, respectively.* From Fig. 27, it is apparent that since at equilibrium, $E_{F1} = E_{F2}$, the total built-in potential is given by

$$\psi_{bi} = \left|\phi_{m1} - \phi_{m2}\right| . \tag{142}$$

The depletion widths and capacitance can be obtained by solving the Poisson equation for the step junction on either side of the interface. One boundary condition is the continuity of electric displacement, that is, $\mathscr{D}_1 = \mathscr{D}_2 = \varepsilon_{s1}\mathscr{E}_1 = \varepsilon_{s2}\mathscr{E}_2$ at the interface. We obtain

$$W_{D1} = \left[\frac{2N_{A2}\varepsilon_{s1}\varepsilon_{s2}(\psi_{bi} - V)}{qN_{D1}(\varepsilon_{s1}N_{D1} + \varepsilon_{s2}N_{A2})}\right]^{1/2} , \tag{143a}$$

$$W_{D2} = \left[\frac{2N_{D1}\varepsilon_{s1}\varepsilon_{s2}(\psi_{bi} - V)}{qN_{A2}(\varepsilon_{s1}N_{D1} + \varepsilon_{s2}N_{A2})}\right]^{1/2} , \tag{143b}$$

and

$$C_D = \left[\frac{qN_{D1}N_{A2}\varepsilon_{s1}\varepsilon_{s2}}{2(\varepsilon_{s1}N_{D1} + \varepsilon_{s2}N_{A2})(\psi_{bi} - V)}\right]^{1/2} . \tag{144}$$

The relative voltage supported in each semiconductor is

$$\frac{\psi_{b1} - V_1}{\psi_{b2} - V_2} = \frac{N_{A2}\varepsilon_{s2}}{N_{D1}\varepsilon_{s1}} \tag{145}$$

where the applied voltage is divided into the two regions $V = V_1 + V_2$. It is apparent that the foregoing expressions will reduce to the expression for the *p-n* junction (homojunction) discussed in Section 2.3, when both sides of the heterojunction become the same materials.

In considering the current flow, the example in Fig. 27b shows that the conduction-band edge E_C increases monotonically while the valence-band edge E_V goes through some peak near the junction. The hole current could become complicated because of the added barrier which might present a bottle-neck in thermionic emission, in series with diffusion. The analysis can be greatly simplified by assuming a graded junction where ΔE_C and ΔE_V become smooth transitions inside the depletion region. With this assumption, the diffusion currents are similar to a regular *p-n* junction but with the appropriate parameters in place. The electron and hole diffusion currents are:

---

* The convention is to list the material with the smaller bandgap as the first symbol.

$$J_n = \frac{qD_{n2}n_{i2}^2}{L_{n2}N_{A2}}\left[\exp\left(\frac{qV}{kT}\right) - 1\right], \tag{146a}$$

$$J_p = \frac{qD_{p1}n_{i1}^2}{L_{p1}N_{D1}}\left[\exp\left(\frac{qV}{kT}\right) - 1\right]. \tag{146b}$$

Note that the band offsets ΔE_C and ΔE_V are not in these equations, and also that each diffusion current component depends on the properties of the receiving side only, as in the case of a homojunction. The total current becomes

$$J = J_n + J_p = \left(\frac{qD_{n2}n_{i2}^2}{L_{n2}N_{A2}} + \frac{qD_{p1}n_{i1}^2}{L_{p1}N_{D1}}\right)\left[\exp\left(\frac{qV}{kT}\right) - 1\right]. \tag{147}$$

Of particular interest is the ratio of the two diffusion currents.

$$\begin{aligned}\frac{J_n}{J_p} &= \frac{L_{p1}D_{n2}N_{D1}n_{i2}^2}{L_{n2}D_{p1}N_{A2}n_{i1}^2} = \frac{L_{p1}D_{n2}N_{D1}N_{C2}N_{V2}\exp(-E_{g2}/kT)}{L_{n2}D_{p1}N_{A2}N_{C1}N_{V1}\exp(-E_{g1}/kT)} \\ &\approx \frac{N_{D1}}{N_{A2}}\exp\left(\frac{-\Delta E_g}{kT}\right) .\end{aligned} \tag{148}$$

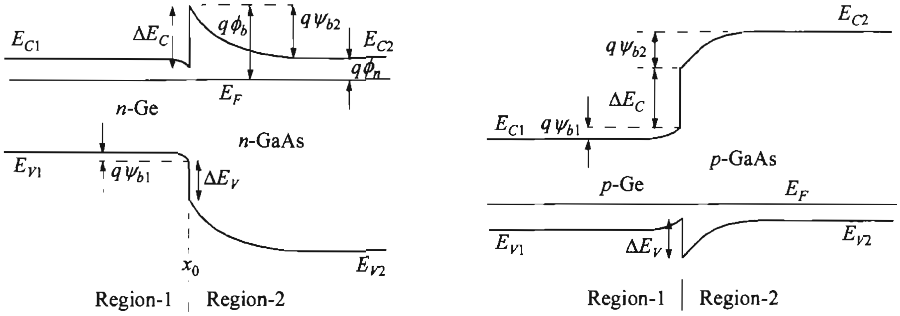
Therefore the injection ratio depends exponentially on the bandgap difference, in addition to their doping ratio. This is critical in designing a bipolar transistor where the injection ratio is directly related to the current gain. The heterojunction bipolar transistor (HBT) uses a wide-bandgap emitter to suppress the base current and will be discussed in more details in Chapter 5.

2.7.2 Isotype Heterojunction

The case of an isotype heterojunction is somewhat different. In an *n-n* heterojunction, since the work function of the wide-bandgap semiconductor is smaller, the energy bands will be bent oppositely to those for the *n-p* case (Fig. 28a).[42] The relation between $(\psi_{b1} - V_1)$ and $(\psi_{b2} - V_2)$ can be found from the boundary condition of continuity of electric displacement ($\mathscr{D} = \varepsilon_s \mathscr{E}$) at the interface. For an accumulation (increase of carriers at the interface) in Region-1 governed by Boltzmann statistics, the electric field at x_0 is given by (for detailed derivation see footnote on p. 84)

$$\mathscr{E}_1(x_0) = \sqrt{\frac{2qN_{D1}}{\varepsilon_{s1}}\left\{\frac{kT}{q}\left[\exp\frac{q(\psi_{b1} - V_1)}{kT} - 1\right] - (\psi_{b1} - V_1)\right\}}. \tag{149}$$

The electric field at the interface for a depletion in Region-2 is given by



(a) **(b)**

Fig. 28 Energy-band diagrams for ideal (a) *n-n* and (b) *p-p* isotype heterojunctions. (After Refs. 40 and 42.)

$$\mathscr{E}_2(x_0) = \sqrt{\frac{2qN_{D2}(\psi_{b2} - V_2)}{\varepsilon_{s2}}}. \tag{150}$$

Equating the electric displacement $\mathscr{D} = \mathscr{E}\varepsilon_s$ of Eqs. 149 and 150 gives a relation between $(\psi_{b1} - V_1)$ and $(\psi_{b2} - V_2)$ that is quite complicated. However, if the ratio $\varepsilon_{s1}N_{D1}/\varepsilon_{s2}N_{D2}$ is of the order of unity and $\psi_{bi}(\equiv \psi_{b1} + \psi_{b2}) \gg kT/q$, we obtain[42]

$$\exp\left[\frac{q(\psi_{b1} - V_1)}{kT}\right] \approx \frac{q}{kT}(\psi_{bi} - V) \tag{151}$$

where V is the total applied voltage and is equal to $(V_1 + V_2)$. Also shown in Fig. 28b is the idealized equilibrium energy-band diagram for *p-p* heterojunctions.

For the carrier transport, because of the potential barrier as shown in Fig. 28a, the conduction mechanism is governed by thermionic emission of majority carriers, electrons in this case (refer to Chapter 3 for details). The current density is given by[42]

$$J = qN_{D2}\sqrt{\frac{kT}{2\pi m_2^*}}\exp\left(\frac{-q\psi_{b2}}{kT}\right)\left[\exp\left(\frac{qV_2}{kT}\right) - \exp\left(\frac{-qV_1}{kT}\right)\right]. \tag{152}$$

Substituting Eq. 151 into Eq. 152 yields the current-voltage relationship:

$$J = \frac{q^2N_{D2}\psi_{bi}}{\sqrt{2\pi m_2^* kT}}\exp\left(\frac{-q\psi_{bi}}{kT}\right)\left(1 - \frac{V}{\psi_{bi}}\right)\left[\exp\left(\frac{qV}{kT}\right) - 1\right]. \tag{153}$$

Since the current is thermionic emission as in a metal-semiconductor contact, the pre-exponential factor is often expressed in terms of the effective Richardson constant A^* and the barrier height ϕ_b. With substitution for A^* and the appropriate expression for N_{D2}, the current equation above becomes

$$
\begin{aligned}
J &= \frac{q\psi_{bi}A^*T}{k}\left(1-\frac{V}{\psi_{bi}}\right)\exp\left(\frac{-q\psi_{bi}}{kT}\right)\exp\left(\frac{-q\phi_b}{kT}\right)\left[\exp\left(\frac{qV}{kT}\right)-1\right] \\
&= J_0\left[\exp\left(\frac{qV}{kT}\right)-1\right] . \qquad (154)
\end{aligned}
$$

This expression is quite different from that for metal-semiconductor contact. The value of J_0 is different [from $A^*T^2\exp(-q\phi_B/kT)$] and so is its temperature dependence. The reverse current never saturates but increases linearly with voltage at large $-V$. In the forward direction, the dependence of J on V can be approximated by an exponential function $J \propto \exp(qV/\eta kT)$.

REFERENCES

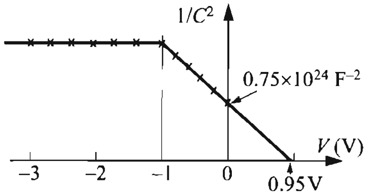
1. W. Shockley, "The Theory of *p-n* Junctions in Semiconductors and *p-n* Junction Transistors," *Bell Syst. Tech. J.*, **28**, 435 (1949);
2. W. Shockley, *Electrons and Holes in Semiconductors*, D. Van Nostrand, Princeton, New Jersey, 1950.
3. C. T. Sah, R. N. Noyce, and W. Shockley, "Carrier Generation and Recombination in *p-n* Junction and *p-n* Junction Characteristics," *Proc. IRE*, **45**, 1228 (1957).
4. J. L. Moll, "The Evolution of the Theory of the Current-Voltage Characteristics of *p-n* Junctions," *Proc. IRE*, **46**, 1076 (1958).
5. C. G. B. Garrett and W. H. Brattain, "Physical Theory of Semiconductor Surfaces," *Phys. Rev.*, **99**, 376 (1955).
6. C. Kittel and H. Kroemer, *Thermal Physics*, 2nd Ed., W. H. Freeman and Co., San Francisco, 1980.
7. W. C. Johnson and P. T. Panousis, "The Influence of Debye Length on the *C-V* Measurement of Doping Profiles," *IEEE Trans. Electron Devices*, **ED-18**, 965 (1971).
8. B. R. Chawla and H. K. Gummel, "Transition Region Capacitance of Diffused *p-n* Junctions," *IEEE Trans. Electron Devices*, **ED-18**, 178 (1971).
9. M. Shur, *Physics of Semiconductor Devices*, Prentice-Hall, Englewood Cliffs, New Jersey, 1990.
10. H. K. Gummel, "Hole-Electron Product of *p-n* Junctions," *Solid-State Electron.*, **10**, 209 (1967).
11. J. L. Moll, *Physics of Semiconductors*, McGraw-Hill, New York, 1964.
12. M. J. O. Strutt, *Semiconductor Devices*, Vol. 1, *Semiconductor and Semiconductor Diodes*, Academic, New York, 1966, Chapter 2.
13. P. J. Lundberg, private communication.
14. S. M. Sze and G. Gibbons, "Avalanche Breakdown Voltages of Abrupt and Linearly Graded *p-n* Junctions in Ge, Si, GaAs, and GaP," *Appl. Phys. Lett.*, **8**, 111 (1966).
15. R. M. Warner, Jr., "Avalanche Breakdown in Silicon Diffused Junctions," *Solid-State Electron.*, **15**, 1303 (1972).
16. M. H. Lee and S. M. Sze, "Orientation Dependence of Breakdown Voltage in GaAs," *Solid-State Electron.*, **23**, 1007 (1980).

17. F. Waldhauser, private communication.
18. S. K. Ghandhi, *Semiconductor Power Devices*, Wiley, New York, 1977.
19. C. R. Crowell and S. M. Sze, "Temperature Dependence of Avalanche Multiplication in Semiconductors," *Appl. Phys. Lett.*, **9**, 242 (1966).
20. C. Y. Chang, S. S. Chiu, and L. P. Hsu, "Temperature Dependence of Breakdown Voltage in Silicon Abrupt *p-n* Junctions," *IEEE Trans. Electron Devices*, **ED-18**, 391 (1971).
21. S. M. Sze and G. Gibbons, "Effect of Junction Curvature on Breakdown Voltages in Semiconductors," *Solid-State Electron.*, **9**, 831 (1966).
22. A. Rusu, O. Pietrareanu, and C. Bulucea, "Reversible Breakdown Voltage Collapse in Silicon Gate-Controlled Diodes," *Solid-State Electron.*, **23**, 473 (1980).
23. A. S. Grove, O. Leistiko, Jr., and W. W. Hooper, "Effect of Surface Fields on the Breakdown Voltage of Planar Silicon *p-n* Junctions," *IEEE Trans. Electron Devices*, **ED-14**, 157 (1967).
24. R. H. Kingston, "Switching Time in Junction Diodes and Junction Transistors," *Proc. IRE*, **42**, 829 (1954).
25. A. Van der Ziel, *Noise in Measurements*, Wiley, New York, 1976.
26. A. Van der Ziel and C. H. Chenette, "Noise in Solid State Devices," in *Advances in Electronics and Electron Physics*, Vol. 46, Academic, New York, 1978.
27. K. K. Ng, *Complete Guide to Semiconductor Devices*, 2nd Ed., Wiley, New York, 2002.
28. J. P. Levin, "Theory of Varistor Electronic Properties", *Crit. Rev. Solid State Sci.*, **5**, 597 (1975).
29. M. H. Norwood and E. Shatz, "Voltage Variable Capacitor Tuning—A Review," *Proc. IEEE*, **56**, 788 (1968).
30. R. A. Moline and G. F. Foxhall, "Ion-Implanted Hyperabrupt Junction Voltage Variable Capacitors," *IEEE Trans. Electron Devices*, **ED-19**, 267 (1972).
31. G. Lucovsky, R. F. Schwarz, and R. B. Emmons, "Transit-Time Considerations in *p-i-n* Diodes," *J. Appl. Phys.*, **35**, 622 (1964).
32. A. G. Milnes, *Semiconductor Devices and Integrated Electronics*, Van Nostrand, New York, 1980
33. W. Shockley, U.S. Patent 2,569,347 (1951).
34. A. I. Gubanov, *Zh. Tekh. Fiz.*, **21**, 304 (1951); *Zh. Eksp. Teor. Fiz.*, **21**, 721 (1951).
35. H. Kroemer, "Theory of a Wide-Gap Emitter for Transistors," *Proc. IRE*, **45**, 1535 (1957).
36. H. C. Casey, Jr., and M. B. Panish, *Heterostructure Lasers*, Academic, New York, 1978.
37. A. G. Milnes and D. L. Feucht, *Heterojunctions and Metal-Semiconductor Junctions*, Academic, New York, 1972.
38. B. L. Sharma and R. K. Purohit, *Semiconductor Heterojunctions*, Pergamon, London, 1974.
39. P. Bhattacharya, Ed., *III-V Quantum Wells and Superlattices*, INSPEC, London, 1996.
40. R. L. Anderson, "Experiments on Ge-GaAs Heterojunctions," *Solid-State Electron.*, **5**, 341 (1962).
41. W. R. Frensley and H. Kroemer, "Theory of the Energy-Band Lineup at an Abrupt Semiconductor Heterojunction," *Phys. Rev. B*, **16**, 2642 (1977).
42. L. L. Chang, "The Conduction Properties of Ge-GaAs$_{1-x}$P$_x$ *n-n* Heterojunctions," *Solid-State Electron.*, **8**, 721 (1965).

PROBLEMS

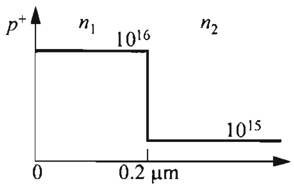
1. A silicon *p-n* junction of 1 cm^2 area consists of a two-sided step junction with an *n*-region of 10^{17} donors/cm^3 and a *p*-region of 2×10^{17} acceptors/cm^3. All donors and acceptors are ionized. Find the built-in potential.

2. The measured depletion capacitance of a p^+-*n* silicon junction (formed in an *n*-type epitaxial layer) is shown. The device area is 10^{-5} cm^2 and the p^+-layer thickness is 0.07 µm. Find the thickness of the epitaxial layer.



3. A silicon *p-n* junction has a linearly graded junction on the *p*-side with an impurity gradient of 10^{19} cm^{-4}, and a uniform doping of 3×10^{14} cm^{-3} on the *n*-side. (a) If the depletion width of the *p*-side is 0.8 µm at zero bias, find the total depletion width, the built-in potential and the maximum field at thermal equilibrium. (b) Plot the impurity and field distribution of this junction.

4. Find the depletion-layer width and the maximum field at thermal equilibration for the p^+-n_1-n_2 structure.



5. (a) A silicon p^+-*n* junction has the following parameters at 300 K: $\tau_p = \tau_g = 10^{-6}$ s, $N_D = 10^{15}$ cm^{-3}. Find the generation current density in the depletion region and the total reverse current density at a bias of 5 V.
   (b) Will there be any significant change of the total reverse current density if τ_p is reduced by a factor of 100 while τ_g remains the same?

6. A p^+-n junction is formed in an *n*-type substrate with $N_D = 10^{15}$ cm^{-3}. If the junction contains 10^{15} cm^{-3} generation-recombination centers located 0.02 eV above the intrinsic Fermi level of silicon with $\sigma_n = \sigma_p = 10^{-15}$ cm^2 ($v_{th} = 10^7$ cm/s), calculate the generation and recombination current at −0.5 V.

7. For a *p-n* junction with the *p*-side doped to 1×10^{17} cm^{-3}, the *n*-side doped to 1×10^{19} cm^{-3}, and a reverse bias of −2 V, calculate the generation-recombination current density, assuming that the effective lifetime is 1×10^{-5} s.

8. Design an abrupt Si p^+-*n* junction diode that has a reverse breakdown voltage of 130 V and has a forward-bias current of 2.2 mA at $V = 0.7$ volt. Assume $\tau_p = 10^{-7}$ s.

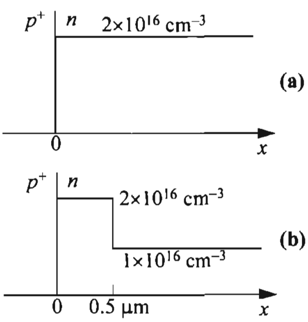
9. (a) Assume $\alpha = \alpha_0(\mathscr{E}/\mathscr{E}_0)^m$ where α_0, $\mathscr{E}_0$, and m are constants. Also assume $\alpha_n = \alpha_p = \alpha$. Derive an expression for the avalanche breakdown voltage of an n^+-*p* junction with a uniform acceptor concentration N_A and a dielectric permittivity ε_s.
   (b) If $\alpha_0 = 10^4$ cm^{-1}, $\mathscr{E}_0 = 4\times10^5$ V/cm, $m = 6$, $N_A = 2\times10^{16}$ cm^{-3} and $\varepsilon_s = 10^{-12}$ F/cm, what is the breakdown voltage?

10. When a silicon p^+-n junction is reverse-biased to 30 V, the depletion-layer capacitance is 1.75 nF/cm^2. If the maximum electric field at avalanche breakdown is 3.1×10^5 V/cm, find the breakdown voltage.

11. A silicon junction diode has a doping profile of p^+-i-n^+-i-n^+ which contains a very narrow n^+-region sandwiched between two i-regions. This narrow region has a doping of 10^{18} cm^{-3} and a width of 10 nm. The first i-region has a thickness of 0.2 μm, and the second i-region is 0.8 μm in thickness. Find the electric field in the second i-region (i.e., in the n^+-i-n^+) when a reverse bias of 20 V is applied to the junction diode.

12. For a silicon one-sided p^+-n-n^+ abrupt junction with a donor concentration of 5×10^{14} cm^{-3}, the maximum field at breakdown is 3×10^5 V/cm. If the thickness of the n-type epitaxial layer is reduced to 5 μm, find the breakdown voltage.

13. For a Si p^+-n one-sided abrupt junction with $N_D = 2\times10^{16}$ cm^{-3}, the breakdown voltage is 32 V (Fig. a). If the doping distribution is modified to Fig. b, find the breakdown voltage.


p^+
n
2×10^{16} cm^{-3}
(a)
0
x
p^+
n
2×10^{16} cm^{-3}
(b)
1×10^{16} cm^{-3}
0
0.5 μm
x

14. Find the value of the electron multiplication factor M_n for a silicon p^+-i-n^+ diode having a reverse bias of 200 V. The corresponding capacitance of the diode is 1.05 nF/cm^2.

15. In an "ideal" silicon n^+-p junction with $N_A = 10^{16}$ cm^{-3}, a minority carrier lifetime of 10^{-8} s, and a mobility of 966 cm^2/V-s, find the stored minority carriers in the neutral p-region of 1 μm, under a forward bias of 1 V.

16. For an ideal abrupt silicon p^+-n junction with $N_D = 10^{15}$ cm^{-3}, find the stored minority carriers (in C/cm^2) in the neutral region when a forward bias of 1 V is applied. Assume the length of the neutral region is 1 μm and the diffusion length of holes is 5 μm. The hole distribution is given by

$$p_n - p_{no} = p_{no}\left[\exp\left(\frac{qV}{kT}\right) - 1\right]\exp\left[\frac{-(x-x_n)}{L_p}\right].$$

17. For a hyperabrupt p^+-n junction varactor, the n-side doping profile is given by $n(x) = Bx^m$ where B is a constant and $m = -3/2$. Derive the express for the differential capacitance.

18. Consider an ideal abrupt heterojunction with a built-in potential of 1.6 V. The impurity concentrations in semiconductor 1 and 2 are 1×10^{16} donors/cm^3 and 3×10^{19} acceptors/cm^3, and dielectric constants are 12 and 13, respectively. Find the electrostatic potential and depletion width in each material for applied voltages of 0.5 V and –5 V.

19. For an n-GaAs/p-Al$_{0.3}$Ga$_{0.7}$As heterojunctlon at room-temperature, $\Delta E_C = 0.21$ eV. (1) What type of heterojunction is this? (2) Based on the Anderson Model, find the total depletion width at thermal equilibrium when both sides have impurity concentration of 5×10^{15} cm^{-3}. (3) Draw the band diagram. [*Hint*: For the bandgap of AlGaAs, refer to Fig. 32 of Chapter 1. The dielectric constant is $(12.9 - 3.12x)$ for Al$_x$Ga$_{1-x}$As. Assume N_C and N_V are the same for Al$_x$Ga$_{1-x}$As with $0 < x < 0.4$.]

20. The alignment of heterojunction between GaAs and $Al_{0.4}Ga_{0.6}As$ is Type-I. The doping concentration is 10^{20} cm^{-3} in $Al_{0.4}Ga_{0.6}As$ and 10^{16} cm^{-3} in GaAs, both doped with carbon. (a) Find the total depletion width under thermal equilibrium condition, assuming the dielectric constant is the same for both semiconductors. (b) Draw the band diagram for $V = 0$.

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

3

Metal-Semiconductor Contacts

3.1 INTRODUCTION
3.2 FORMATION OF BARRIER
3.3 CURRENT TRANSPORT PROCESSES
3.4 MEASUREMENT OF BARRIER HEIGHT
3.5 DEVICE STRUCTURES
3.6 OHMIC CONTACT

3.1 INTRODUCTION

The earliest systematic investigation on metal-semiconductor rectifying systems is generally attributed to Braun, who in 1874 noted the dependence of the total resistance of a point contact on the polarity of the applied voltage and on the detailed surface conditions.[1] The point-contact rectifier in various forms found practical applications beginning in 1904.[2] In 1931, Wilson formulated the transport theory of semiconductors based on the band theory of solids.[3] This theory was then applied to metal-semiconductor contacts. In 1938, Schottky suggested that the potential barrier could arise from stable space charges in the semiconductor alone without the presence of a chemical layer.[4] The model arising from this consideration is known as the Schottky barrier. In 1938, Mott also devised a more appropriate theoretical model for swept-out metal-semiconductor contacts that is known as the Mott barrier.[5] These models were further enhanced by Bethe in 1942 to become the thermionic-emission model which accurately describes the electrical behavior.[6] The basic theory, the historical development, and the device technology of rectifying metal-semiconductor contacts can be found in References 7–11.

Because of their importance in direct current and microwave applications and as intricate parts of other semiconductor devices, metal-semiconductor contacts have been studied extensively. Specifically, they have been used as photodetectors, solar cells, as the gate electrode of the MESFET, etc. Most importantly, the metal contact on heavily doped semiconductor forms an ohmic contact that is required for every semiconductor device in order to pass current in and out of the device.

3.2 FORMATION OF BARRIER

When metal makes contact with a semiconductor, a barrier is formed at the metal-semiconductor interface. This barrier is responsible for controlling the current conduction as well as its capacitance behavior. In this section, we consider the basic energy-band diagrams leading to the formation of the barrier height and some effects that can modify the value of this barrier.

3.2.1 Ideal Condition

We will first consider the ideal case without surface states and other anomalies. Figure 1a shows the electronic energy relations of a high work-function metal and an *n*-type semiconductor which are not in contact and are in separate systems. If the two are allowed to communicate with each other, for example by an external wire connection, charge will flow from the semiconductor to the metal and thermal equilibrium is established as a single system. The Fermi levels on both sides will line up. Relative to the Fermi level in the metal, the Fermi level in the semiconductor is lowered by an amount equal to the difference between the two work functions.

The work function is the energy difference between the vacuum level and the Fermi level. This quantity is denoted by $q\phi_m$ for the metal, and is equal to $q(\chi + \phi_n)$ in the semiconductor, where $q\chi$ is the electron affinity measured from the bottom of the conduction band E_C to the vacuum level, and $q\phi_n$ is the energy difference between E_C

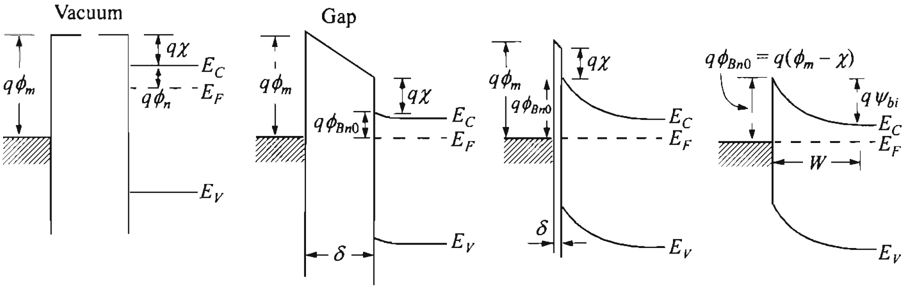


Fig. 1 Energy-band diagrams of metal-semiconductor contacts. Metal and semiconductor (a) in separated systems, and (b) connected into one system. As the gap δ (c) is reduced and (d) becomes zero. (After Ref. 7.)

and the Fermi level. The potential difference between the two work functions $\phi_m - (\chi + \phi_n)$ is called the contact potential. As the gap distance δ decreases, the electric field in the gap increases and an increasing negative charge is built up at the metal surface. An equal and opposite charge (positive) must exist in the semiconductor depletion region. The potential variation within the depletion layer is similar to that in one side of a *p-n* junction. When δ is small enough to be comparable to the interatomic distances, the gap becomes transparent to electrons, and we obtain the limiting case, as shown on the far right (Fig. 1d). It is clear that the limiting value of the barrier height $q\phi_{Bn0}$ is given by

$$q\phi_{Bn0} = q(\phi_m - \chi). \tag{1}$$

The barrier height is simply the difference between the metal work function and the electron affinity of the semiconductor. Conversely, for an ideal contact between a metal and a *p*-type semiconductor, the barrier height $q\phi_{Bp0}$ is given by

$$q\phi_{Bp0} = E_g - q(\phi_m - \chi). \tag{2}$$

Thus, for any given semiconductor and metal combination, the sum of the barrier heights on *n*-type and *p*-type substrates is expected to be equal to the bandgap, or

$$q(\phi_{Bn0} + \phi_{Bp0}) = E_g. \tag{3}$$

In practice, however, simple expressions for the barrier heights as given by Eqs. 1 and 2 are never realized experimentally. The electron affinities of semiconductors and the work functions of metals have been established. For metals, $q\phi_m$ is of the order of a few electron volts (2 – 6 eV). The values of $q\phi_m$ are generally very sensitive to surface contamination. The most reliable values for clean surfaces are given in Fig. 2. The main deviations of experimental barrier heights from the ideal condition are: (1) an unavoidable interface layer, $\delta \neq 0$ as in Fig. 1c, and (2) the presence of interface states. Furthermore, the barrier height can be modified due to image-force lowering. These effects will be discussed in the following sections.

3.2.2 Depletion Layer

The depletion layer of a metal-semiconductor contact is similar to that of the one-sided abrupt (e.g., p^+-n) junction. It is clear from the discussion above that when a metal is brought into intimate contact with a semiconductor, the conduction and valence bands of the semiconductor at the surface are brought into a definite energy relationship with the Fermi level in the metal. Once this relationship is established, it serves as a boundary condition to the solution of the Poisson equation in the semiconductor, which proceeds in exactly the same manner as in a *p-n* junction. The energy-band diagrams for metals on both *n*-type and *p*-type materials are shown, under different biasing conditions, in Fig. 3.

For contacts on *n*-type semiconductors, under the abrupt approximation that $\rho \approx qN_D$ for $x < W_D$, $\rho \approx 0$ and $\mathscr{E} \approx 0$ for $x > W_D$, where W_D is the depletion width, we obtain

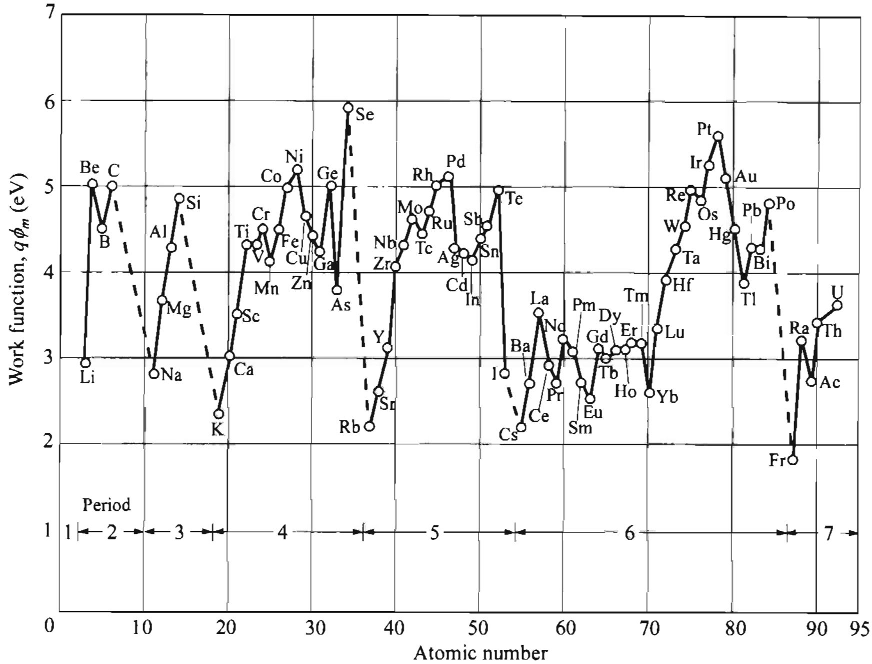


Fig. 2 Metal work function for a clean metal surface in a vacuum versus atomic number. Note the periodic nature of the increase and decrease of the work functions within each group. (After Ref. 12.)

$$W_D = \sqrt{\frac{2\varepsilon_s}{qN_D}\left(\psi_{bi} - V - \frac{kT}{q}\right)} \tag{4}$$

$$|\mathscr{E}(x)| = \frac{qN_D}{\varepsilon_s}(W_D - x) = \mathscr{E}_m - \frac{qN_D x}{\varepsilon_s} \tag{5}$$

$$E_C(x) = q\phi_{Bn} - \frac{q^2 N_D}{\varepsilon_s}\left(W_D x - \frac{x^2}{2}\right) \tag{6}$$

where the term kT/q arises from the contribution of the majority-carrier distribution tail (electrons in n-side, see the footnote on p. 84) and $\mathscr{E}_m$ is the maximum field strength which occurs at $x = 0$:

$$\mathscr{E}_m = \mathscr{E}(x = 0) = \sqrt{\frac{2qN_D}{\varepsilon_s}\left(\psi_{bi} - V - \frac{kT}{q}\right)} = \frac{2[\psi_{bi} - V - (kT/q)]}{W_D}. \tag{7}$$

The space charge Q_{sc} per unit area of the semiconductor and the depletion-layer capacitance C_D per unit area are given by

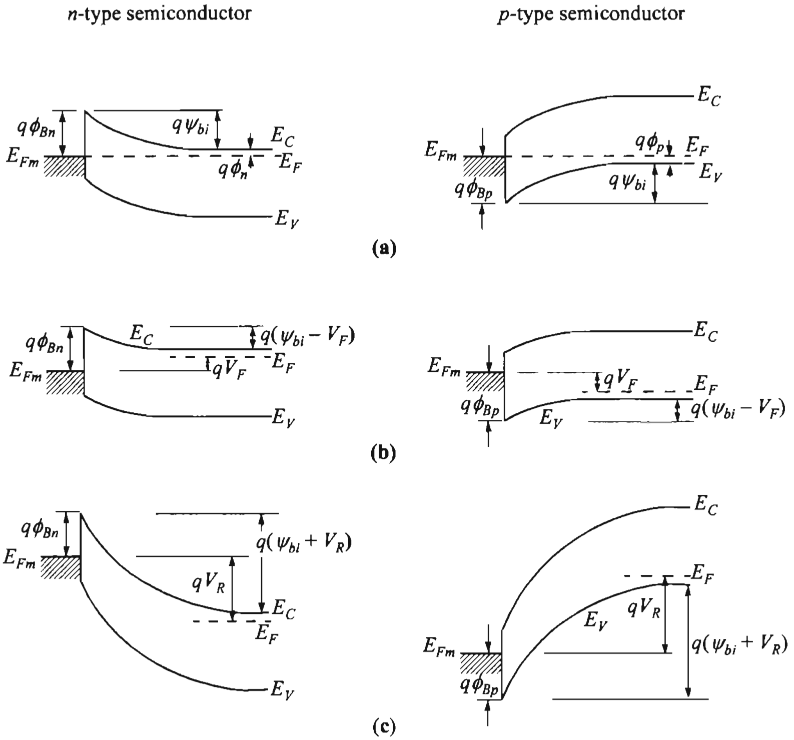

n-type semiconductor
p-type semiconductor
$q\phi_{Bn}$
E_{Fm}
$q\psi_{bi}$
E_C
E_F
$q\phi_n$
E_V
$q\phi_p$
$q\phi_{Bp}$
(a)
$q(\psi_{bi} - V_F)$
qV_F
(b)
$q(\psi_{bi} + V_R)$
qV_R
(c)

Fig. 3 Energy-band diagrams of metal on *n*-type (left) and on *p*-type (right) semiconductors under different biasing conditions. (a) Thermal equilibrium. (b) Forward bias. (c) Reverse bias.

$$Q_{sc} = qN_DW_D = \sqrt{2q\varepsilon_sN_D\left(\psi_{bi} - V - \frac{kT}{q}\right)} \tag{8}$$

$$C_D \equiv \frac{\varepsilon_s}{W_D} = \sqrt{\frac{q\varepsilon_sN_D}{2[\psi_{bi} - V - (kT/q)]}}. \tag{9}$$

Equation 9 can be written in the form

$$\frac{1}{C_D^2} = \frac{2[\psi_{bi} - V - (kT/q)]}{q\varepsilon_sN_D} \tag{10}$$

or

$$N_D = \frac{2}{q\varepsilon_s}\left[-\frac{1}{d(1/C_D^2)/dV}\right]. \tag{11}$$

If N_D is constant throughout the depletion region, one should obtain a straight line by plotting $1/C_D^2$ versus voltage. If N_D is not a constant, the differential capacitance method can be used to determine the doping profile from Eq. 11, similar to the case of a one-sided *p-n* junction as discussed in Section 2.2.1.

The *C-V* measurement can also be used to study deep impurity levels. Figure 4 shows a semiconductor with one shallow donor level and one deep donor level.[13] While all the shallow donors above the Fermi level will be ionized, only deep impurities near the surface are above the Fermi level and ionized, giving a higher effective doping concentration near the interface. In a *C-V* measurement where a small ac signal is superimposed on the dc bias, there will be a frequency dependence on capacitance since the deep impurities can only follow slow signals, i.e. dN_T/dV is absent at high frequencies. Comparing *C-V* measurements at various frequencies can reveal the properties of these deep-level impurities.

3.2.3 Interface States

The barrier heights of metal-semiconductor systems are, in general, determined by both the metal work function and the interface states. A general expression of the barrier height can be obtained on the basis of the following two assumptions:[14] (1) with intimate contact between the metal and the semiconductor, and with an interfa-

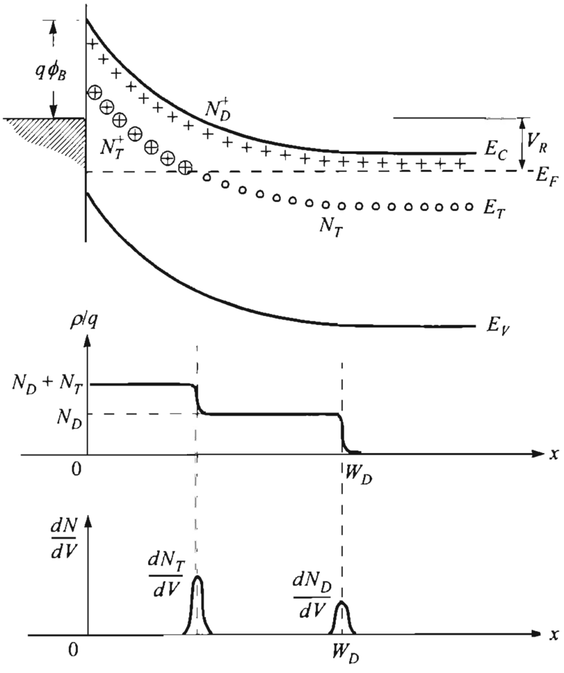
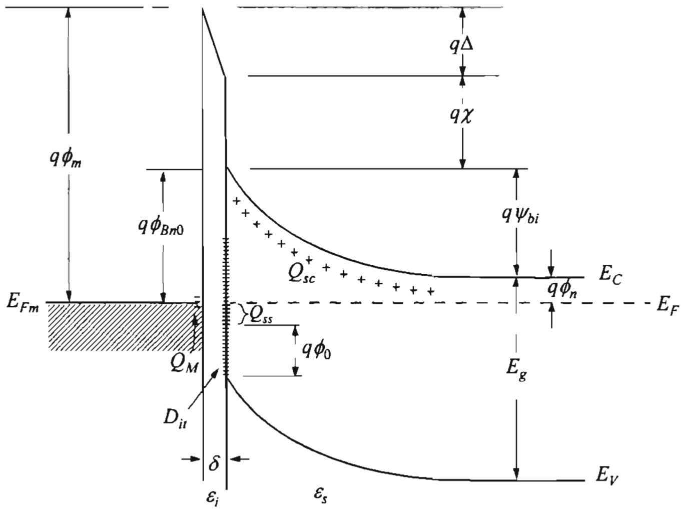


Fig. 4 Semiconductor with one shallow donor level and one deep donor level. N_D and N_T are the shallow donor and deep donor concentration, respectively. (After Ref. 13.)

cial layer of atomic dimensions, this layer will be transparent to electrons but can withstand potential across it, and (2) the interface states per unit area per energy at the interface are a property of the semiconductor surface and are independent of the metal. A more detailed energy-band diagram of a practical metal-*n*-semiconductor contact is shown in Fig. 5. The various quantities used in the derivation that follows are defined in this figure. The first quantity of interest is the energy level $q\phi_0$ above E_V at the semiconductor surface. It is called the neutral level above which the states are of acceptor type (neutral when empty, negatively charged when full) and below which the states are of donor type (neutral when full of electrons, positively charged when empty). Consequently, when the Fermi level at the surface coincides with this neutral level, the net interface-trap charge is zero.[15] This energy level also tends to pin the semiconductor Fermi level at the surface before the metal contact was formed.



ϕ_m = Work function of metal
ϕ_{Bn0} = Barrier height (without image-force lowering)
ϕ_0 = Neutral level (above E_V) of interface states
Δ = Potential across interfacial layer
χ = Electron affinity of semiconductor
ψ_{bi} = Built-in potential
δ = Thickness of interfacial layer
Q_{sc} = Space-charge density in semiconductor
Q_{ss} = Interface-trap charge
Q_M = Surface-charge density on metal
D_{it} = Interface-trap density
ε_i = Permittivity of interfacial layer (vacuum)
ε_s = Permittivity of semiconductor

Fig. 5 Detailed energy-band diagram of a metal-*n*-semiconductor contact with an interfacial layer (vacuum) of the order of atomic distance. (After Ref. 14.)

The second quantity is $q\phi_{Bn0}$, the barrier height of the metal-semiconductor contact; it is this barrier that must be surmounted by electrons flowing from the metal into the semiconductor. The interfacial layer will be assumed to have a thickness of a few angstroms and will therefore be essentially transparent to electrons.

We consider a semiconductor with acceptor interface traps (since in this particular example E_F is above the neutral level) whose density is D_{it} states/cm^2-eV, and is a constant over the energy range from $q\phi_0 + E_V$ to the Fermi level. The interface-trap charge density on the semiconductor Q_{ss} is therefore negative and is given by

$$Q_{ss} = -qD_{it}(E_g - q\phi_0 - q\phi_{Bn0}) \qquad \text{C/cm}^2. \tag{12}$$

The quantity in parentheses is simply the energy difference between the Fermi level at the surface and the neutral level. The interface-trap density D_{it} times this quantity yields the number of surface states above the neutral level that are full.

The space charge that forms in the depletion layer of the semiconductor at thermal equilibrium is given as

$$Q_{sc} = qN_D W_D = \sqrt{2q\varepsilon_s N_D\left(\phi_{Bn0} - \phi_n - \frac{kT}{q}\right)}. \tag{13}$$

The total equivalent surface charge density on the semiconductor surface is given by the sum of Eqs. 12 and 13. In the absence of any space-charge effects in the interfacial layer, an exactly equal and opposite charge, Q_M (C/cm^2), develops on the metal surface. For thin interfacial layers such space-charge effects are negligible and Q_M can be written as

$$Q_M = -(Q_{ss} + Q_{sc}). \tag{14}$$

The potential Δ across the interfacial layer can be obtained by applying Gauss' law to the surface charge on the metal and semiconductor:

$$\Delta = -\frac{\delta Q_M}{\varepsilon_i} \tag{15}$$

where ε_i is the permittivity of the interfacial layer and δ its thickness. Another relation for Δ can be obtained by inspection of the energy-band diagram of Fig. 5:

$$\Delta = \phi_m - (\chi + \phi_{Bn0}). \tag{16}$$

This relation results from the fact that the Fermi level must be constant throughout this system at thermal equilibrium.

If Δ is eliminated from Eqs. 15 and 16, and Eq. 14 is used to substitute for Q_M, we obtain

$$\phi_m - \chi - \phi_{Bn0} = \sqrt{\frac{2q\varepsilon_s N_D \delta^2}{\varepsilon_i^2}\left(\phi_{Bn0} - \phi_n - \frac{kT}{q}\right)} - \frac{qD_{it}\delta}{\varepsilon_i}(E_g - q\phi_0 - q\phi_{Bn0}). \tag{17}$$

Equation 17 can now be solved for ϕ_{Bn0}. We introduce the quantities

$$c_1 \equiv \frac{2q\varepsilon_s N_D \delta^2}{\varepsilon_i^2}, \tag{18}$$

$$c_2 \equiv \frac{\varepsilon_i}{\varepsilon_i + q^2 \delta D_{it}} \tag{19}$$

which contain all the interfacial properties. Equation 18 can be used to calculate c_1 if values of δ and ε_i are estimated. For vacuum-cleaved or well-cleaned semiconductor substrates the interfacial layer will have a thickness of atomic dimensions (i.e., 4 or 5 Å). The permittivity of such a thin layer can be well approximated by the free-space value and since this approximation represents a lower limit for ε_i, it leads to an overestimation of c_2. For $\varepsilon_s \approx 10\varepsilon_0$, $\varepsilon_i = \varepsilon_0$, and $N_D < 10^{18}$ cm^{-3}, c_1 is small, of the order of 0.01 V and the square-root term in Eq. 17 is estimated to be less than 0.1 V. Neglecting this square-root term, Eq. 17 reduces to

$$\phi_{Bn0} = c_2(\phi_m - \chi) + (1 - c_2)\left(\frac{E_g}{q} - \phi_0\right) \equiv c_2 \phi_m + c_3 . \tag{20}$$

With known c_2 and c_3 from experiments of varying ϕ_m, the interfacial properties are given by

$$\phi_0 = \frac{E_g}{q} - \frac{c_2 \chi + c_3}{1 - c_2} , \tag{21}$$

$$D_{it} = \frac{(1 - c_2)\varepsilon_i}{c_2 \delta q^2} . \tag{22}$$

Using the previous assumptions for δ and ε_i, we obtain $D_{it} \approx 1.1\times10^{13}(1 - c_2)/c_2$ states/cm^2-eV.

There are two limiting cases which can be obtained directly from Eq. 20:

1. When $D_{it} \rightarrow \infty$, then $c_2 \rightarrow 0$ and

$$q\phi_{Bn0} = E_g - q\phi_0 . \tag{23}$$

In this case the Fermi level at the interface is *pinned* by the surface states at the value $q\phi_0$ above the valence band. The barrier height is independent of the metal work function and is determined entirely by the surface properties of the semiconductor.

2. When $D_{it} \rightarrow 0$, then $c_2 \rightarrow 1$ and

$$q\phi_{Bn0} = q(\phi_m - \chi) . \tag{24}$$

This equation for the barrier height of an ideal Schottky barrier where surface-state effects are neglected, is identical to Eq. 1.

The experimental results of the metal-n-silicon system are shown in Fig. 6a. A least-square straight-line fit to the data yields

$$q\phi_{Bn0} = 0.27 q\phi_m - 0.52 . \tag{25}$$

Comparing this expression with Eq. 20 ($c_2 = 0.27$, $c_3 = -0.52$) and using Eqs. 21 and 22, we obtain $q\phi_0 = 0.33$ eV, and $D_{it} = 4\times10^{13}$ states/cm^2-eV. Similar results are obtained for GaAs, GaP, and CdS, which are shown in Fig. 6b and listed in Table 1.

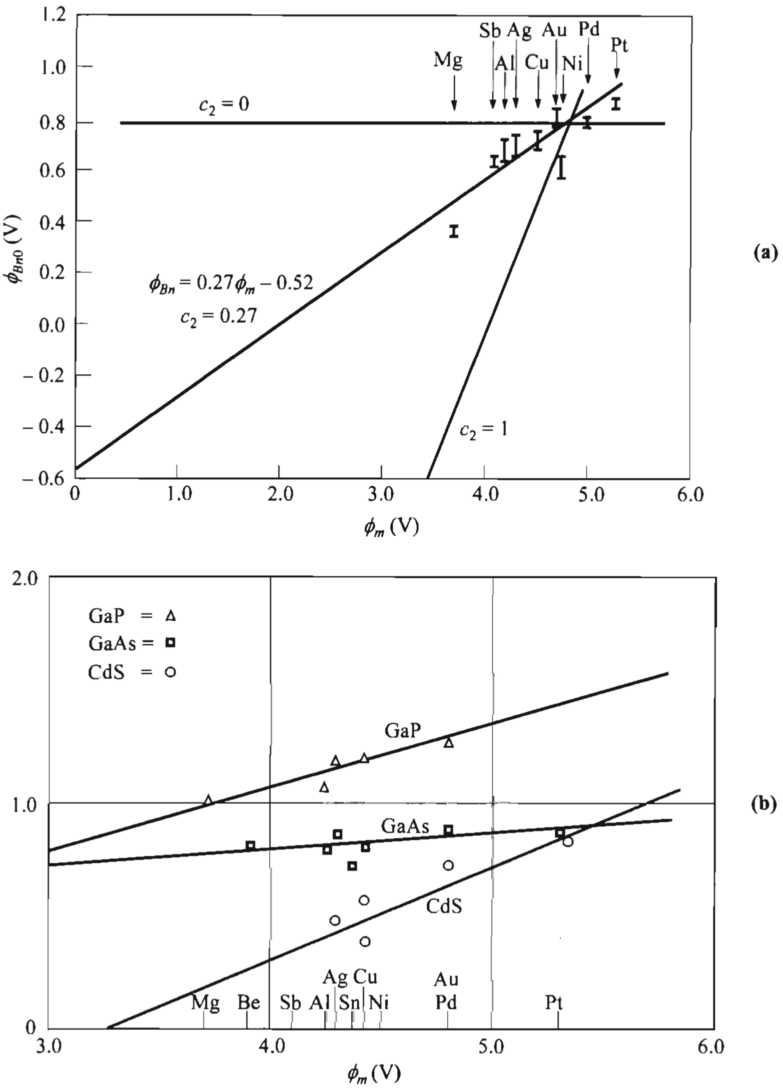


Fig. 6 Experimental barrier heights for different metals on *n*-type (a) silicon and (b) GaAs, GaP, and CdS. (After Ref. 14.)

It should be pointed out that in spite of nonideal factors such as interface states, the relationship of Eq. 3, that the sum of barrier heights on *n*- and *p*-type substrates equals the energy gap of the semiconductor, is still generally valid.

We note that the values of $q\phi_0$ for Si, GaAs, and GaP are very close to one-third of the bandgap. Similar results are obtained for other semiconductors.[16] This fact indicates that most covalent semiconductor surfaces have a high peak density of

Table 1 Summary of Barrier Height Data and Calculations of Interface Properties for Si, GaAs, GaP, and CdS (After Ref. 14)

Semi-conductor	c_2	c_3 (V)	χ (V)	D_{it} (10^{13} /eV-cm^2)	$q\phi_0$ (eV)	$q\phi_0/E_g$
Si	0.27±0.05	−0.52±0.22	4.05	2.7±0.7	0.30±0.36	0.27
GaAs	0.07±0.05	0.51±0.24	4.07	12.5±10.0	0.53±0.33	0.38
GaP	0.27±0.03	0.02±0.13	4.0	2.7±0.4	0.66±0.2	0.294
CdS	0.38±0.16	−1.17±0.77	4.8	1.6±1.1	1.5±1.5	0.6

surface states or defects near the neutral level and that the neutral level is about one-third of the bandgap from the valence-band edge. The theoretical calculation by Pugh[17] for ⟨111⟩ diamond indeed gives a narrow band of surface states slightly below the center of the forbidden gap. It is thus expected that a similar situation may exist for other semiconductors.

For III-V compounds, extensive measurements using photoemission spectroscopy indicate that the Schottky-barrier formation is due mainly to defects generated near the interface by deposition of the metal.[18] It has been shown that on a few compound semiconductors such as GaAs, GaSb, and InP, the surface Fermi-level positions obtained from a number of metals are pinned at an energy level quite independent of the metal.[19] This pinning of surface Fermi level can explain the fact that for most III-V compounds, the barrier height is essentially independent of metal work function.

For ionic semiconductors such as CdS and ZnS, the barrier height generally depends strongly on the metal and a correlation has been found between interface behavior and the electronegativity. The electronegativity X_M is defined as the power of an atom in a molecule to attract electrons to itself. Figure 7 shows Pauling's electronegativity scale. Note that the periodicity is similar to that for the work function (Fig. 2).

Figure 8a shows a plot of the barrier height versus the electronegativity of metals deposited on Si, GaSe, and SiO$_2$. From the plot we define the slope as an index of interface behavior:

$$S \equiv \frac{d\phi_{Bn0}}{dX_M}. \tag{26}$$

Note the comparison of S to c_2 ($= d\phi_{Bn0}/d\phi_m$). We can also plot the index S as a function of the electronegativity difference (ionicity ΔX) of the semiconductors, shown in Fig. 8b. The electronegativity difference is defined as the difference in the Pauling electronegativities between the cation and the anion of the semiconductor. Note a sharp transition from the covalent semiconductors (such as GaAs with $\Delta X = 0.4$) to ionic semiconductors (such as AlN with $\Delta X = 1.5$). For semiconductors with $\Delta X < 1$, the index S is small, indicating that the barrier height is only weakly dependent on metal electronegativity (or the work function). On the other hand, for $\Delta X > 1$, the

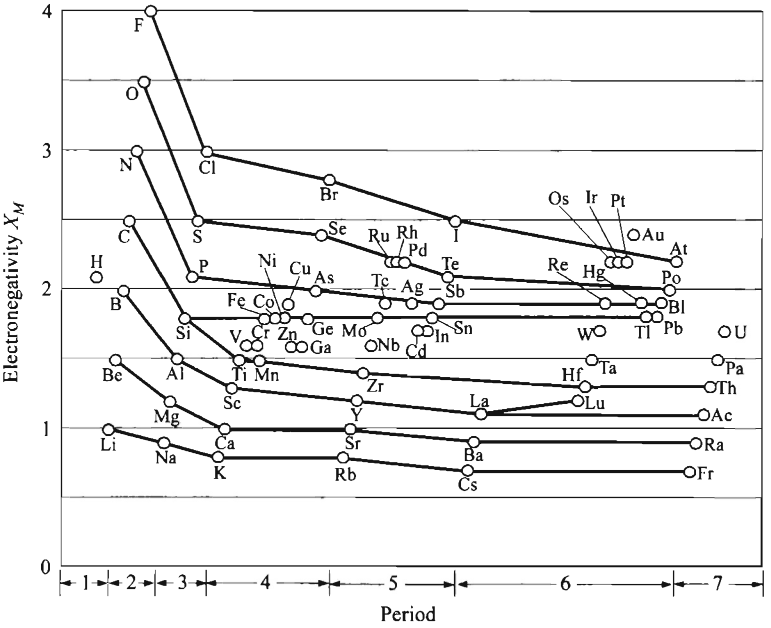


Fig. 7 Pauling's electronegativity scale. Note the trend of increasing electronegativity within each group. (After Ref. 20.)

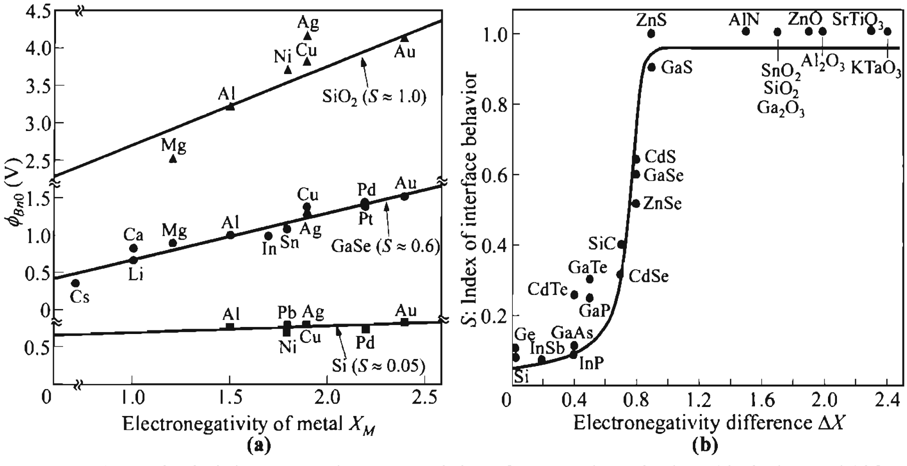


Fig. 8 (a) Barrier height versus electronegativity of metals deposited on Si, GaSe, and SiO_2. (b) Index of interface behavior S as a function of the electronegativity difference of the semiconductors. (After Ref. 21.)

index S approaches 1, and the barrier height is strongly dependent on the metal electronegativity (or the work function).

For technological applications in silicon integrated circuits, an important class of Schottky barrier contacts has been developed in which a chemical reaction between the metal and the underlying silicon is induced to form silicides.[22] The formation of metal silicides by solid-solid metallurgical reaction provides more reliable and reproducible Schottky barriers, because the interface chemical reactions are well defined and can be maintained under good control. It is thought that since the silicide interfacial properties depends on the eutectic temperature, there should be a correlation between the barrier height and the eutectic temperature. Figure 9 shows such an empirical fit for the barrier heights on n-type silicon of transition-metal silicides plotted against the eutectic temperature of the silicides. Similar correlation had been observed when barrier heights are plotted against the heat of formation of silicides.[24]

3.2.4 Image-Force Lowering

The image-force lowering, also known as the Schottky effect or Schottky-barrier lowering, is the image-force-induced lowering of the barrier energy for charge carrier emission, in the presence of an electric field. Consider a metal-vacuum system first. The minimum energy necessary for an electron to escape into vacuum from an initial energy at the Fermi level is the work function $q\phi_m$ as shown in Fig. 10. When an electron is at a distance x from the metal, a positive charge will be induced on the metal surface. The force of attraction between the electron and the induced positive charge is equivalent to the force that would exist between the electron and an equal positive

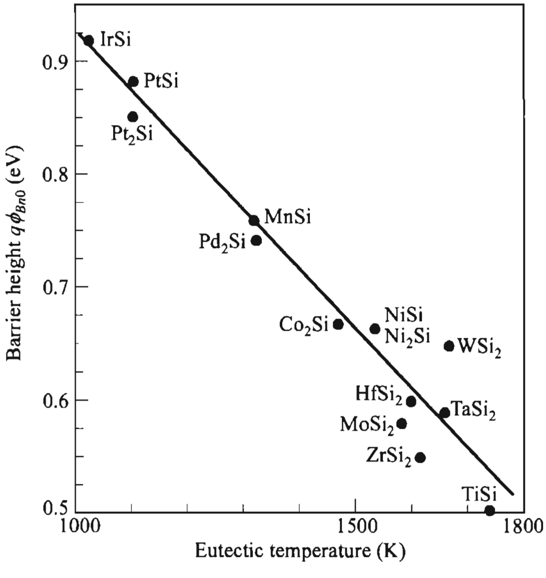


Fig. 9 Correlation of barrier height of transition-metal silicides with their eutectic temperature. (After Ref. 23.)

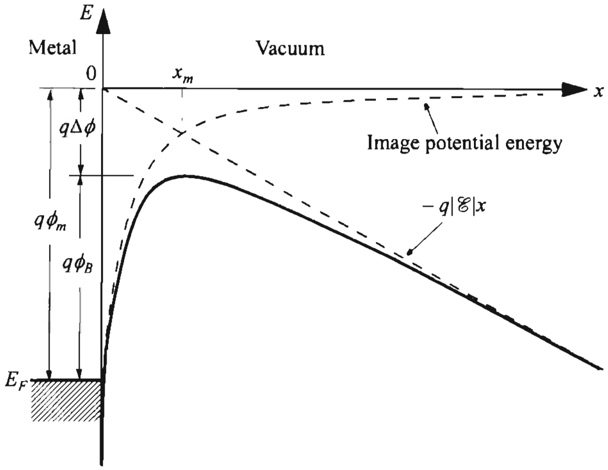


Fig. 10 Energy-band diagram between a metal surface and a vacuum. The metal work function is $q\phi_m$. The effective barrier is lowered when an electric field is applied to the surface. The lowering is due to the combined effects of the field and the image force.

charge located at $-x$. This positive charge is referred to as the image charge. The attractive force toward the metal, called the image force, is given by

$$F = \frac{-q^2}{4\pi\varepsilon_0(2x)^2} = \frac{-q^2}{16\pi\varepsilon_0 x^2} \tag{27}$$

where ε_0 is the permittivity of free space. The work done to an electron in the course of its transfer from infinity to the point x is given by

$$E(x) = \int_{\infty}^{x} F dx = \frac{-q^2}{16\pi\varepsilon_0 x}. \tag{28}$$

This energy corresponds to the potential energy of an electron placed at a distance x from the metal surface, shown in Fig. 10, and is measured downwards from the x-axis. When an external field $\mathscr{E}$ is applied (in this example in the $-x$ direction), the total potential energy PE as a function of distance is given by the sum

$$PE(x) = -\frac{q^2}{16\pi\varepsilon_0 x} - q|\mathscr{E}|x. \tag{29}$$

This equation has a maximum value. The image-force lowering $\Delta\phi$ and the location of the lowering x_m (as shown in Fig. 10), are given by the condition $d(PE)/dx = 0$, or

$$x_m = \sqrt{\frac{q}{16\pi\varepsilon_0|\mathscr{E}|}} \tag{30}$$

$$\Delta\phi = \sqrt{\frac{q|\mathscr{E}|}{4\pi\varepsilon_0}} = 2|\mathscr{E}|x_m. \tag{31}$$

From Eqs. 30 and 31 we obtain $\Delta\phi = 0.12$ V and $x_m = 6$ nm for $\mathscr{E} = 10^5$ V/cm; and $\Delta\phi = 1.2$ V and $x_m = 1$ nm for $\mathscr{E} = 10^7$ V/cm. Thus at high fields the Schottky barrier is considerably lowered, and the effective metal work function for thermionic emission ($q\phi_B$) is reduced.

These results can be applied to metal-semiconductor systems. However, the field should be replaced by the appropriate field at the interface, and the free-space permittivity ε_0 should be replaced by an appropriate permittivity ε_s characterizing the semiconductor medium, that is,

$$\Delta\phi = \sqrt{\frac{q\mathscr{E}_m}{4\pi\varepsilon_s}}. \tag{32}$$

Note that inside a device such as metal-semiconductor contact, the field is not zero even without bias due to the built-in potential. Because of the larger values of ε_s in a metal-semiconductor system, the barrier lowering is smaller than that in a corresponding metal-vacuum system. For example, for $\varepsilon_s = 12\varepsilon_0$, $\Delta\phi$ as obtained from Eq. 32 is only 0.035 V for $\mathscr{E} = 10^5$ V/cm and even smaller for smaller fields. Also a typical value for x_m is calculated to be less than 5 nm. Although the barrier lowering is small, it does have a profound effect on current transport processes in metal-semiconductor systems. These are considered in Section 3.3.

In a practical Schottky-barrier diode, the electric field is not constant with distance, and the maximum value at the surface based on the depletion approximation can be used,

$$\mathscr{E}_m = \sqrt{\frac{2qN|\psi_s|}{\varepsilon_s}}, \tag{33}$$

where the surface potential ψ_s (on n-type substrate) is

$$|\psi_s| = \phi_{Bn0} - \phi_n + V_R. \tag{34}$$

Substituting $\mathscr{E}_m$ into Eq. 32 gives

$$\Delta\phi = \sqrt{\frac{q\mathscr{E}_m}{4\pi\varepsilon_s}} = \left[\frac{q^3N|\psi_s|}{8\pi^2\varepsilon_s^3}\right]^{1/4}. \tag{35}$$

Figure 11 shows the energy diagram incorporating the Schottky effect for a metal on n-type semiconductor under different biasing conditions. Note that for forward bias (V > 0), the field and the image force are smaller and the barrier height $q\phi_{Bn0} - q\Delta\phi_F$ is slightly larger than the barrier height at zero bias of

$$q\phi_{Bn} = q\phi_{Bn0} - q\Delta\phi. \tag{36}$$

For reverse bias ($V_R > 0$), the barrier height $q\phi_{Bn0} - q\Delta\phi_R$ is slightly smaller. In effect, the barrier height becomes bias dependant.

The value ε_s may also be different from the semiconductor static permittivity. If during the emission process, the electron transit time from the metal-semiconductor interface to the barrier maximum x_m is shorter than the dielectric relaxation time, the semiconductor medium does not have enough time to be polarized, and smaller per-

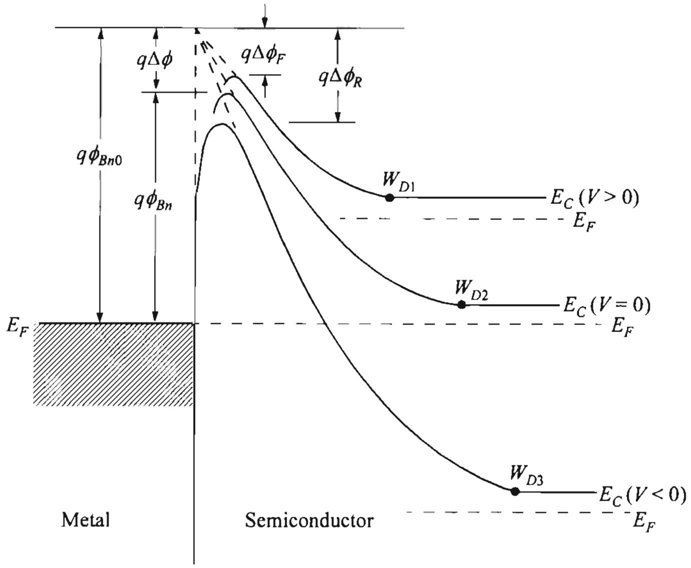


Fig. 11 Energy-band diagram incorporating the Schottky effect for a metal *n*-type semiconductor contact under different biasing conditions. The intrinsic barrier height is $q\phi_{Bn0}$. The barrier height at thermal equilibrium is $q\phi_{Bn}$. The barrier lowerings under forward and reverse bias are $\Delta\phi_F$ and $\Delta\phi_R$ respectively. (After Ref. 10.)

mittivity than the static value is expected. It will be shown, however, that for Si the appropriate permittivities are about the same as their corresponding static values.

The dielectric constant ($K_s = \varepsilon_s/\varepsilon_0$) in gold-silicon barriers has been obtained from photoelectric measurements, which will be discussed in Section 3.4.4. The experimental results are shown in Fig. 12, where the measured barrier lowering is plotted as a function of the square root of the maximum electric field.[25] From Eq. 35 the image-force dielectric constant is determined to be 12 ± 0.5. For $\varepsilon_s/\varepsilon_0 = 12$, the distance x_m varies between 1 and 5 nm for the field range shown in Fig. 12. Assuming a carrier velocity of the order of 10^7 cm/s, the transit time for these distances should be between $1–5\times10^{-14}$ s. The image-force dielectric constant should thus be comparable to the value of approximately 12 for electromagnetic radiation of roughly these periods (wavelengths between 3 and 15 μm).[26] The dielectric constant of bulk silicon is essentially constant (11.7) from dc to $\lambda = 1$ μm, therefore the lattice has time to polarize while the electron is traversing the depletion layer. The photoelectric measurements and data deduced from the optical constants are in excellent agreement. For Ge and GaAs, the dependence of the optical dielectric constant on wavelength is similar to that of Si. The image-force permittivities of these semiconductors in the foregoing field range are thus expected to be approximately the same as the corresponding static bulk values.

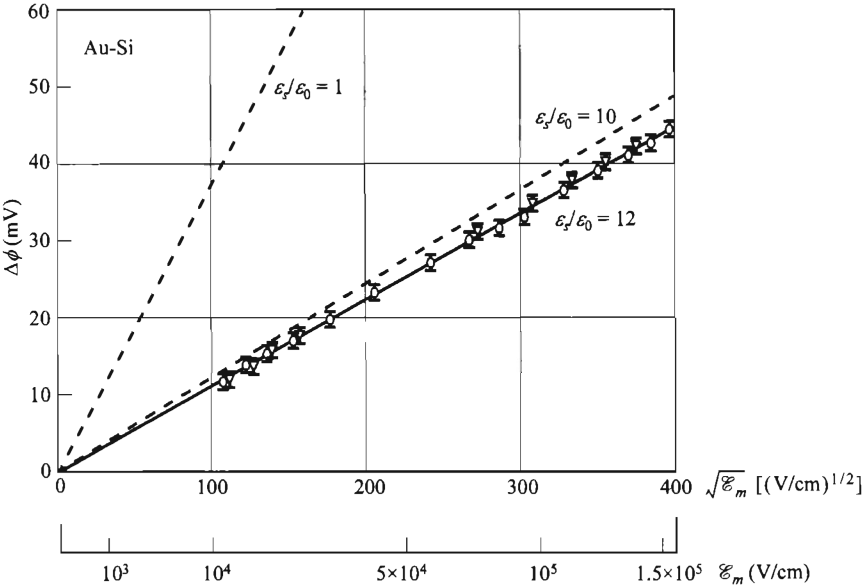


Fig. 12 Measurement of barrier lowering as a function of the electric field in a Au-Si diode. (After Ref. 25.)

3.2.5 Barrier-Height Adjustment

For an ideal Schottky barrier, the barrier height is determined primarily by the characters of the metal and the metal-semiconductor interface properties and is nearly independent of the doping. Usual Schottky barriers on a given semiconductor (e.g., n- or p-type Si) therefore give a finite number of choices for barrier height. However, by introducing a thin layer ($\approx$ 10 nm or less) of controllable number of dopants on a semiconductor surface (e.g., by ion implantation), the effective barrier height for a given metal-semiconductor contact can be varied.[27–29] This approach is particularly useful in order to select a metal having the most desirable metallurgical properties required for reliable device operation and at the same time to be able to adjust the effective barrier height between this metal and the semiconductor in a controlled manner.

Figure 13a shows the idealized controlled barrier contacts with a thin n^+-layer or a thin p^+-layer on an n-type substrate for barrier reduction or barrier increase, respectively. Consider the reduction of barrier first. The field distribution in Fig. 13b is given by

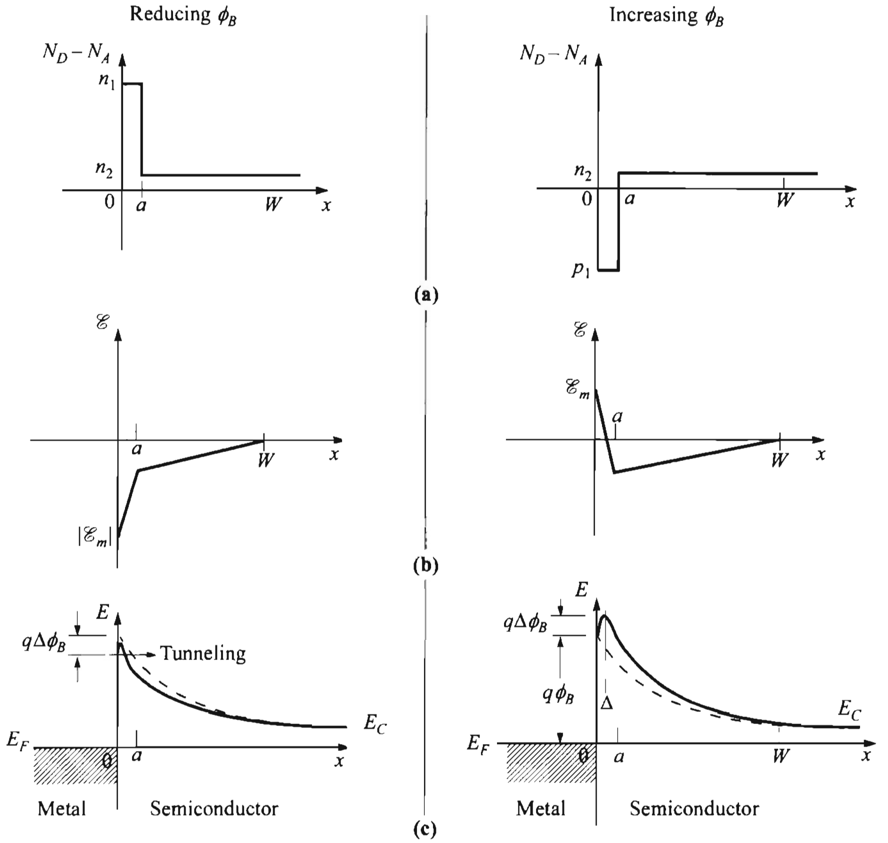

Reducing ϕ_B
Increasing ϕ_B
$N_D - N_A$
n_1
n_2
p_1
0
a
W
x
(a)
$\mathscr{E}$
$|\mathscr{E}_m|$
$\mathscr{E}_m$
(b)
E
$q\Delta\phi_B$
$q\phi_B$
Δ
Tunneling
E_F
E_C
Metal
Semiconductor
(c)

Fig. 13 Idealized controlled barrier contacts with a thin n^+-layer or a thin p^+-layer on an n-type substrate for barrier reduction (left) or barrier increase (right), respectively. Dashed lines indicate original barrier with uniform doping.

$$\begin{aligned} \mathscr{E} &= -|\mathscr{E}_m| + \frac{qn_1x}{\varepsilon_s} && \text{for} \quad 0<x<a \\ &= -\frac{qn_2}{\varepsilon_s}(W-x) && \text{for} \quad a<x<W \end{aligned} \tag{37}$$

where $\mathscr{E}_m$ is the maximum electric field at the metal-semiconductor interface, and is given by

$$|\mathscr{E}_m| = \frac{q}{\varepsilon_s}[n_1a + n_2(W-a)]. \tag{38}$$

The image-force lowering due to $\mathscr{E}_m$ is given by Eq. 35. For Si and GaAs Schottky barriers with n_2 of the order of 10^{16} cm^{-3} or less, the zero-bias value of $n_2(W-a)$ is about 10^{11} cm^{-2}. Therefore, if n_1a is made sufficiently larger than 10^{11} cm^{-2}, Eqs. 38 and 35 can be reduced to

$$|\mathscr{E}_m| \approx \frac{qn_1a}{\varepsilon_s}, \tag{39}$$

$$\Delta\phi \approx \frac{q}{\varepsilon_s}\sqrt{\frac{n_1a}{4\pi}}. \tag{40}$$

For $n_1a = 10^{12}$ and 10^{13} cm^{-2}, the corresponding lowerings are 0.045 and 0.14 V, respectively.

Although the image-force lowering contributes to the barrier reduction, generally the tunneling effect is more significant. For $n_1a = 10^{13}$ cm^{-2}, the maximum field from Eq. 39 is 1.6×10^6 V/cm, which is the zero-bias field of a Au-Si Schottky diode with a doping of 10^{19} cm^{-3}. The increased saturation current density due to tunneling for such a diode is about 10^{-3} A/cm^2, corresponding to an effective barrier height of 0.6 V (see discussion later on current vs. barrier height), a reduction of 0.2 V from the 0.8 V barrier of the original Au-Si diode. The calculated effective barrier height as a function of $\mathscr{E}_m$ is shown in Fig. 14 for Si and GaAs barriers. By increasing the maximum field from 10^5 V/cm to 10^6 V/cm, one generally can reduce the effective barrier by 0.2 V in Si and over 0.3 V in GaAs.

For a given application, the parameters n_1 and a should be properly chosen so that in the forward direction the larger Schottky-barrier lowering and the added tunneling current will not substantially degrade the ideality factor η. And in the reverse direction, they will not cause large leakage current in the required bias range.

If opposite doping is formed in the thin semiconductor layer at the interface, the effective barrier can be increased. As indicated in Fig. 13a, if the n^+-region is replaced

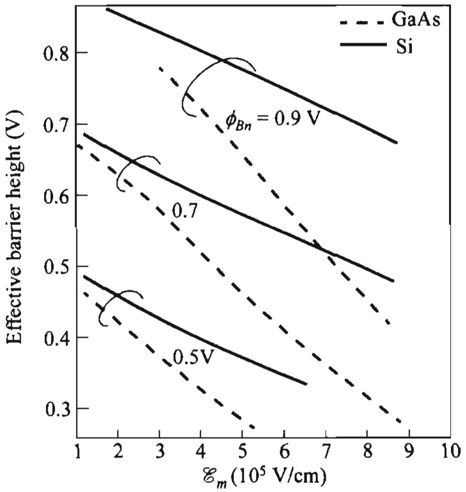


Fig. 14 Calculated reduced effective barrier height from tunneling for Si and GaAs metal-semiconductor contacts. (After Ref. 30.)

by p^+-region, it can be shown that the energy-band profile will be $q\phi_B$ at $x = 0$ and reach a maximum at $x = \Delta$, where

$$\Delta = \frac{1}{p_1}[ap_1 - (W-a)n_2]. \tag{41}$$

The effective barrier height occurs at $x = \Delta$ and is given by

$$\phi'_B = \phi_B + \mathscr{E}_m\Delta - \frac{qp_1\Delta^2}{2\varepsilon_s}. \tag{42}$$

Equation 42 approaches $(\phi_B + qp_1a^2/2\varepsilon_s)$ if $p_1 \gg n_2$ and $ap_1 \gg Wn_2$. Therefore, as the product ap_1 increases, the effective barrier height will increase accordingly.

Figure 15 shows the measured results of Ni-Si diodes with shallow antimony implantation on the surface. As the implant dose increases, the effective barrier height decreases for n-type substrates and increases for p-type substrates.

3.3 CURRENT TRANSPORT PROCESSES

The current transport in metal-semiconductor contacts is due mainly to majority carriers, in contrast to p-n junctions where the minority carriers are responsible. Figure 16 shows five basic transport processes under forward bias (the inverse processes occur under reverse bias).[8] These five processes are (1) emission of electrons from the semiconductor over the potential barrier into the metal [the dominant process for Schottky diodes with moderately doped semiconductors (e.g., Si with $N_D \leq 10^{17}$ cm^{-3}) operated at moderate temperatures (e.g., 300 K)], (2) quantum-mechanical tunneling of electrons through the barrier (important for heavily doped semiconductors and responsible for most ohmic contacts), (3) recombination in the space-charge region [identical to the recombination process in a p-n junction (refer to

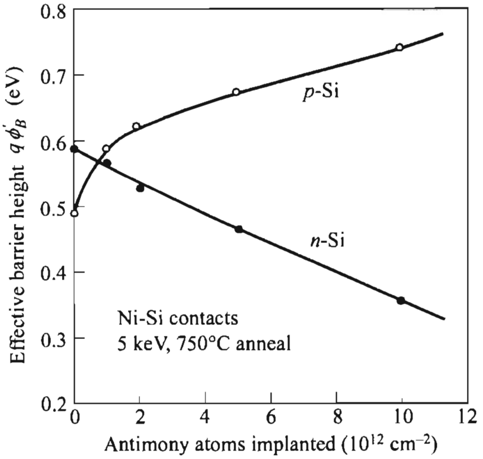


Fig. 15 Effective barrier height for holes in p-type substrates and for electrons in n-type substrates as a function of the implanted antimony dose. (After Ref. 30.)

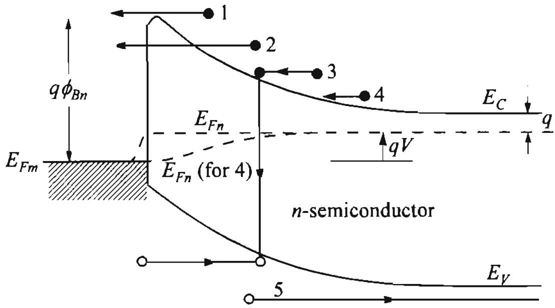


Fig. 16 Five basic transport processes under forward bias. (1) Thermionic emission. (2) Tunneling. (3) Recombination. (4) Diffusion of electrons. (5) Diffusion of holes.

chapter 2)], (4) diffusion of electrons in the depletion region, and (5) holes injected from the metal that diffuse into the semiconductor (equivalent to recombination in the neutral region). In addition, we may have edge leakage current due to a high electric field at the metal-contact periphery or interface current due to traps at the metal-semiconductor interface. Various methods have been used to improve the interface quality, and many device structures have been proposed to reduce or eliminate the edge leakage current (see Section 3.5).

For common high-mobility semiconductors (e.g., Si and GaAs) the transport can be adequately described by this thermionic-emission theory. We shall also consider the diffusion theory applicable to low-mobility semiconductors and a generalized thermionic-emission-diffusion theory that is a synthesis of the preceding two theories.

Schottky diode behavior is to some extent electrically similar to a one-sided abrupt *p-n* junction, and yet the Schottky diode can be operated as a majority-carrier device with inherent fast response. Thus, the terminal functions of a *p-n* junction diode can general be performed by a Schottky diode with one exception as a charge-storage diode. This is because the charge-storage time in a majority-carrier device is extremely small. Another difference is the larger current density in a Schottky diode due to the smaller built-in potential as well as the nature of thermionic emission compared to diffusion. This results in a much smaller forward voltage drop. By the same token, the disadvantage is the larger reverse current in the Schottky diode and a lower breakdown voltage.

3.3.1 Thermionic-Emission Theory

The thermionic-emission theory by Bethe[6] is derived from the assumptions that (1) the barrier height $q\phi_{Bn}$ is much larger than kT, (2) thermal equilibrium is established at the plane that determines emission, and (3) the existence of a net current flow does not affect this equilibrium so that one can superimpose two current fluxes—one from metal to semiconductor, the other from semiconductor to metal, each with a different quasi Fermi level. If thermionic emission is the limiting mechanism, then E_{Fn} is flat throughout the deletion region (Fig. 16). Because of these assumptions, the shape of the barrier profile is immaterial and the current flow depends solely on the barrier height. The current density from the semiconductor to the metal $J_{s\rightarrow m}$ is then given by the concentration of electrons with energies sufficient to overcome the potential barrier and traversing in the x-direction:

$$J_{s\rightarrow m} = \int_{E_{Fn}+q\phi_{Bn}}^{\infty} qv_x dn \tag{43}$$

where $E_{Fn} + q\phi_{Bn}$ is the minimum energy required for thermionic emission into the metal, and v_x is the carrier velocity in the direction of transport. The electron density in an incremental energy range is given by

$$\begin{aligned} dn &= N(E)F(E)dE \\ &\approx \frac{4\pi(2m^*)^{3/2}}{h^3}\sqrt{E-E_C}\exp\left(-\frac{E-E_C+q\phi_n}{kT}\right)dE \end{aligned} \tag{44}$$

where $N(E)$ and $F(E)$ are the density of states and the distribution function, respectively.

If we postulate that all the energy of electrons in the conduction band is kinetic energy, then

$$E - E_C = \frac{1}{2}m^*v^2 \tag{45}$$

$$dE = m^*vdv \tag{46}$$

$$\sqrt{E-E_C} = v\sqrt{\frac{m^*}{2}}. \tag{47}$$

Substituting Eqs. 45–47 into Eq. 44 gives

$$dn \approx 2\left(\frac{m^*}{h}\right)^3 \exp\left(-\frac{q\phi_n}{kT}\right)\exp\left(-\frac{m^*v^2}{2kT}\right)(4\pi v^2 dv) . \tag{48}$$

Equation 48 gives the number of electrons per unit volume that have velocities between v and $v + dv$, distributed over all directions. If the velocity is resolved into its components along the axes with the x-axis parallel to the transport direction, we have

$$v^2 = v_x^2 + v_y^2 + v_z^2 . \tag{49}$$

With the transformation $4\pi v^2 dv = dv_x dv_y dv_z$, we obtain from Eqs. 43, 48, and 49

$$\begin{aligned} J_{s\rightarrow m} &= 2q\left(\frac{m^*}{h}\right)^3 \exp\left(-\frac{q\phi_n}{kT}\right)\int_{v_{0x}}^{\infty} v_x \exp\left(-\frac{m^*v_x^2}{2kT}\right)dv_x \\ &\qquad \int_{-\infty}^{\infty}\exp\left(-\frac{m^*v_y^2}{2kT}\right)dv_y \int_{-\infty}^{\infty}\exp\left(-\frac{m^*v_z^2}{2kT}\right)dv_z \\ &= \left(\frac{4\pi q m^* k^2}{h^3}\right)T^2 \exp\left(-\frac{q\phi_n}{kT}\right)\exp\left(-\frac{m^*v_{0x}^2}{2kT}\right) . \end{aligned} \tag{50}$$

The velocity v_{0x} is the minimum velocity required in the x-direction to surmount the barrier and is given by

$$\frac{1}{2}m^*v_{0x}^2 = q(\psi_{bi} - V). \tag{51}$$

Substituting Eq. 51 into Eq. 50 yields

$$\begin{aligned} J_{s \to m} &= \left(\frac{4\pi q m^* k^2}{h^3}\right) T^2 \exp\left(-\frac{q\phi_{Bn}}{kT}\right) \exp\left(\frac{qV}{kT}\right) \\ &= A^* T^2 \exp\left(-\frac{q\phi_{Bn}}{kT}\right) \exp\left(\frac{qV}{kT}\right) , \end{aligned} \tag{52}$$

and

$$A^* = \frac{4\pi q m^* k^2}{h^3} \tag{53}$$

is the effective Richardson constant for thermionic emission, neglecting the effects of optical-phonon scattering and quantum mechanical reflection (see Section 3.3.3). For free electrons ($m^* = m_0$) the Richardson constant A is 120 A/cm^2-K^2. Note that when the image-force lowering is considered, the barrier height ϕ_{Bn} in Eq. 52 is reduced by $\Delta\phi$.

For semiconductors with isotropic effective mass in the lowest minimum of the conduction band such as n-type GaAs, A^*/A simply is equal to m^*/m_0. For multiple-valley semiconductors the appropriate Richardson constant associated with a single energy minimum is given by[31]

$$\frac{A_1^*}{A} = \frac{1}{m_0}\sqrt{l_1^2 m_y^* m_z^* + l_2^2 m_z^* m_x^* + l_3^2 m_x^* m_y^*} \tag{54}$$

where l_1, l_2, and l_3 are the direction cosines of the normal to the emitting plane relative to the principal axes of the ellipsoid, and m_x^*, m_y^*, and m_z^* are the components of the effective mass tensor.

For Si the conduction band minima occur in the $\langle 100 \rangle$-directions and $m_l^* = 0.98m_0$, $m_t^* = 0.19m_0$. The minimum value of A^* occurs for the $\langle 100 \rangle$-directions:

$$\left(\frac{A^*}{A}\right)_{n\text{-Si}\langle 100 \rangle} = \frac{2m_t^*}{m_0} + \frac{4\sqrt{m_l^* m_t^*}}{m_0} = 2.1 . \tag{55}$$

In the $\langle 111 \rangle$-directions all minima contribute equally to the current, yielding the maximum A^*:

$$\left(\frac{A^*}{A}\right)_{n\text{-Si}\langle 111 \rangle} = \frac{6}{m_0}\sqrt{\frac{(m_t^*)^2 + 2m_l^* m_t^*}{3}} = 2.2 . \tag{56}$$

For holes in Si and GaAs the two energy maxima at $\boldsymbol{k} = 0$ give rise to approximately isotropic current flow from both the light and heavy holes. Adding the currents due to these carriers, we obtain

$$\left(\frac{A^*}{A}\right)_{p\text{-type}} = \frac{m^*_{lh} + m^*_{hh}}{m_0}. \tag{57}$$

Table 2 gives a summary of the values of A^*/A for Si and GaAs.

Since the barrier height for electrons moving from the metal into the semiconductor remains the same under bias, the current flowing into the semiconductor is thus unaffected by the applied voltage. It must therefore be equal to the current flowing from the semiconductor into the metal when thermal equilibrium prevails (i.e., when $V = 0$). This corresponding current density is obtained from Eq. 52 by setting $V = 0$,

$$J_{m\rightarrow s} = -A^*T^2\exp\left(-\frac{q\phi_{Bn}}{kT}\right). \tag{58}$$

The total current density is given by the sum of Eqs. 52 and 58.

$$\begin{aligned} J_n &= \left[A^*T^2\exp\left(-\frac{q\phi_{Bn}}{kT}\right)\right]\left[\exp\left(\frac{qV}{kT}\right) - 1\right] \\ &= J_{TE}\left[\exp\left(\frac{qV}{kT}\right) - 1\right] \end{aligned} \tag{59}$$

where

$$J_{TE} \equiv A^*T^2\exp\left(-\frac{q\phi_{Bn}}{kT}\right). \tag{60}$$

Equation 59 is similar to the transport equation for *p-n* junctions. However, the expressions for the saturation current densities are quite different.

An alternative approach to derive the thermionic-emission current is the following.[8] Without decomposing the velocity components, only electrons with energy above the barrier will contribute to the forward current. This number of electrons above the barrier is given by

$$n = N_C\exp\left[\frac{-q(\phi_{Bn} - V)}{kT}\right]. \tag{61}$$

It is known that for a Maxwellian distribution of velocities, the current from random motion of carriers across a plane is given by

$$J = nq\frac{v_{ave}}{4} \tag{62}$$

where v_{ave} is the average thermal velocity,

Table 2 Values of A^*/A (After Ref. 31)

Semiconductor	Si	GaAs	
p-type	0.66	0.62	
n-type ⟨100⟩	2.1	0.063 (low field)	0.55 (high field)
n-type ⟨111⟩	2.2	”	”

$$v_{ave} = \sqrt{\frac{8kT}{\pi m^*}} \; . \tag{63}$$

Substitution of Eqs. 61 and 63 into Eq. 62 gives

$$J = \frac{4(kT)^2 q \pi m^*}{h^3} \exp\left[\frac{-q(\phi_{Bn} - V)}{kT}\right] \tag{64}$$

which is identical to Eq. 52.

3.3.2 Diffusion Theory

The diffusion theory by Schottky[4] is derived from the assumptions that (1) the barrier height is much larger than kT, (2) the effect of electron collisions within the depletion region, i.e., diffusion, is included, (3) the carrier concentrations at $x = 0$ and $x = W_D$ are unaffected by the current flow (i.e., they have their equilibrium values), and (4) the impurity concentration of the semiconductor is nondegenerate.

Since the current in the depletion region depends on the local field and the concentration gradient, we must use the current density equation:

$$\begin{aligned} J_x = J_n &= q\left(n\mu_n \mathscr{E} + D_n \frac{dn}{dx}\right) \\ &= qD_n\left(\frac{n}{kT}\frac{dE_C}{dx} + \frac{dn}{dx}\right) . \end{aligned} \tag{65}$$

Under the steady-state condition, the current density is independent of x, and Eq. 65 can be integrated using $\exp[E_C(x)/kT]$ as an integrating factor. We then have

$$J_n \int_0^{W_D} \exp\left[\frac{E_C(x)}{kT}\right] dx = qD_n \left\{ n(x) \exp\left[\frac{E_C(x)}{kT}\right] \right\} \Bigg|_0^{W_D} \tag{66}$$

and the boundary conditions using $E_{Fm} = 0$ as the reference (see Fig. 16 but ignore image force for diffusion):

$$E_C(0) = q\phi_{Bn} , \tag{67}$$

$$E_C(W_D) = q(\phi_n + V) , \tag{68}$$

$$n(0) = N_C \exp\left[-\frac{E_C(0) - E_{Fn}(0)}{kT}\right] = N_C \exp\left(-\frac{q\phi_{Bn}}{kT}\right) , \tag{69}$$

$$n(W_D) = N_D = N_C \exp\left(-\frac{q\phi_n}{kT}\right) . \tag{70}$$

Substituting Eqs. 67–70 into Eq. 66 yields

$$J_n = qN_C D_n \left[\exp\left(\frac{qV}{kT}\right) - 1\right] \Bigg/ \int_0^{W_D} \exp\left[\frac{E_C(x)}{kT}\right] dx . \tag{71}$$

For Schottky barriers, neglecting image-force effect, the potential distribution is given by Eq. 6. Substituting this expression for $E_C(x)$ into Eq. 71 and expressing W_D in terms of $\psi_{bi} + V$ leads to

$$J_n \approx \frac{q^2 D_n N_C}{kT}\sqrt{\frac{2qN_D(\psi_{bi}-V)}{\varepsilon_s}}\exp\left(-\frac{q\phi_{Bn}}{kT}\right)\left[\exp\left(\frac{qV}{kT}\right)-1\right]$$
$$\approx q\mu_n N_C \mathscr{E}_m \exp\left(-\frac{q\phi_{Bn}}{kT}\right)\left[\exp\left(\frac{qV}{kT}\right)-1\right] = J_D\left[\exp\left(\frac{qV}{kT}\right)-1\right]. \qquad (72)$$

The current density expressions of the diffusion and thermionic-emission theories, Eqs. 59 and 72, are basically very similar. However, the saturation current density for the diffusion theory J_D is dependent on the bias and is less sensitive to temperature compared to the saturation current density of the thermionic-emission theory J_{TH}.

3.3.3 Thermionic-Emission-Diffusion Theory

A synthesis of the thermionic-emission and diffusion approaches described above has been proposed by Crowell and Sze.[32] This approach is derived from the boundary condition of a thermionic recombination velocity v_R near the metal-semiconductor interface.

Since the diffusion of carriers is strongly affected by the potential configuration in the region through which the diffusion occurs, we consider the electron potential energy [or $E_C(x)$] versus distance incorporating the Schottky lowering effect as shown in Fig. 17. We consider the case where the barrier height is large enough that the charge density between the metal surface and $x = W_D$ is essentially that of the ionized donors (i.e., depletion approximation). As drawn, the applied voltage V between the

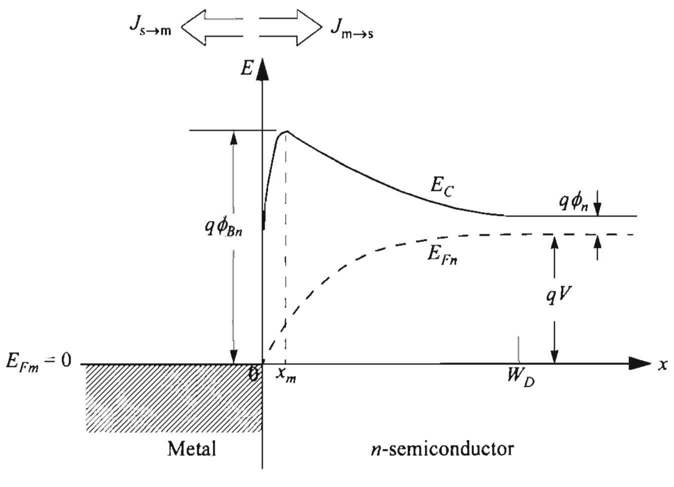


Fig. 17 Energy-band diagram incorporating the Schottky effect to show the derivations of thermionic-emission-diffusion theory and tunneling current.

metal and the semiconductor bulk would give rise to a flow of electrons toward the metal. The electron quasi-Fermi level E_{Fn} in the barrier is also shown schematically as a function of distance. Throughout the region between x_m and W_D,

$$J = n\mu_n \frac{dE_{Fn}}{dx} \tag{73}$$

where the electron density at any point x is given by

$$n = N_C \exp\left(-\frac{E_C - E_{Fn}}{kT}\right). \tag{74}$$

We will assume that the region between x_m and W_D is isothermal and that the electron temperature T is equal to the lattice temperature.

If the portion of the barrier between x_m and the interface ($x = 0$) acts as a sink for electrons, we can describe the current flow in terms of an effective recombination velocity υ_R at the potential energy maximum x_m:

$$J = q(n_m - n_0)\upsilon_R \tag{75}$$

where n_m is the electron density at x_m when the current is flowing,

$$n_m = N_C \exp\left[\frac{E_{Fn}(x_m) - E_C(x_m)}{kT}\right] = N_C \exp\left[\frac{E_{Fn}(x_m) - q\phi_{Bn}}{kT}\right]. \tag{76}$$

n_0 is a quasi-equilibrium electron density at x_m, the density that would occur if it were possible to reach equilibrium without altering the magnitude or position of the potential energy maximum, i.e., $E_{Fn}(x_m) = E_{Fm}$

$$n_0 = N_C \exp\left(-\frac{q\phi_{Bn}}{kT}\right). \tag{77}$$

Another boundary condition, taking $E_{Fm} = 0$ as reference, is

$$E_{Fn}(W_D) = qV. \tag{78}$$

If n is eliminated from Eqs. 73 and 74 and the resulting expression for E_{Fn} is integrated between x_m and W_D,

$$\exp\left[\frac{E_{Fn}(x_m)}{kT}\right] - \exp\left(\frac{qV}{kT}\right) = \frac{-J}{\mu_n N_C kT}\int_{x_m}^{W_D} \exp\left(\frac{E_C}{kT}\right) dx . \tag{79}$$

Then from Eqs. 75 and 79, $E_{Fn}(x_m)$ can be solved as

$$\exp\left[\frac{E_{Fn}(x_m)}{kT}\right] = \frac{\upsilon_D \exp(qV/kT) + \upsilon_R}{\upsilon_D + \upsilon_R} \tag{80}$$

where

$$\upsilon_D \equiv D_n \exp\left(\frac{q\phi_{Bn}}{kT}\right) \bigg/ \int_{x_m}^{W_D} \exp\left[\frac{E_C}{kT}\right] dx \tag{81}$$

is an effective diffusion velocity associated with the transport of electrons from the edge of the depletion layer W_D to the potential energy maximum x_m. Substituting Eq. 80 into Eq. 75 gives the end result of the thermionic-emission-diffusion theory

$$J_{TED} = \frac{qN_C v_R}{1+(v_R/v_D)} \exp\left(-\frac{q\phi_{Bn}}{kT}\right)\left[\exp\left(\frac{qV}{kT}\right) - 1\right]. \tag{82}$$

In this equation, the relative values of v_R and v_D determines the relative contribution of thermionic emission versus diffusion. The parameter v_D can be evaluated as the Dawson's integral and can be approximated by $v_D \approx \mu_n \mathscr{E}_m$ in this case of depletion region.[8] If the electron distribution is Maxwellian for $x \geq x_m$, and if no electrons return from the metal other than those associated with the current density qn_0v_R, the semiconductor acts as a thermionic emitter. Then v_R is the thermal velocity given by

$$\begin{aligned} v_R &= \int_0^\infty v_x \exp\left(\frac{-m^* v_x^2}{2kT}\right) dv_x \Bigg/ \int_{-\infty}^\infty \exp\left(\frac{-m^* v_x^2}{2kT}\right) dv_x \\ &= \sqrt{\frac{kT}{2m^*\pi}} = \frac{A^* T^2}{qN_C} \end{aligned} \tag{83}$$

where A^* is the effective Richardson constant, as shown in Table 2. At 300 K, v_R is 5.2×10^6 and 1.0×10^7 cm/s for $\langle 111 \rangle$ n-type Si and n-type GaAs respectively. It can be seen that if $v_D \gg v_R$, the pre-exponential term in Eq. 82 is dominated by v_R and the thermionic-emission theory applies ($J_{TED} = J_{TE}$). If, however, $v_D \ll v_R$, the diffusion process is the limiting factor ($J_{TED} = J_D$).

In summary, Eq. 82 gives a result that is a synthesis of Schottky's diffusion theory and Bethe's thermionic-emission theory, and it predicts currents in essential agreement with the thermionic-emission theory if $\mu\mathscr{E}(x_m) > v_R$. The latter criterion is more rigorous than Bethe's condition $\mathscr{E}(x_m) > kT/q\lambda$, where λ is the carrier mean free path.

In the preceding section a recombination velocity v_R associated with thermionic emission was introduced as a boundary condition to describe the collecting action of the metal in a Schottky barrier. In many cases an appreciable probability exists that an electron which crosses the potential energy maximum will be back-scattered by electron optical-phonon scattering.[33,34] As a first approximation the probability of electron emission over the potential maximum can be given by $f_p = \exp(-x_m/\lambda)$. In addition, the electron energy distribution can be further distorted from a Maxwellian distribution because of quantum-mechanical reflection of electrons by the Schottky barrier, and also because of tunneling of electrons through the barrier.[35,36] The ratio f_Q of the total current flow, considering the quantum-mechanical tunneling and reflection, to the current flow neglecting these effects depends strongly on the electric field and the electron energy measured from the potential maximum.

The complete expression of the J-V characteristics taking into account f_p and f_Q is thus

$$J = A^{**} T^2 \exp\left(-\frac{q\phi_{Bn}}{kT}\right)\left[\exp\left(\frac{qV}{kT}\right) - 1\right] \tag{84}$$

where

$$A^{**} = \frac{f_p f_Q A^*}{1 + (f_p f_Q v_R / v_D)}. \tag{85}$$

The impacts of these effects are reflected in the reduced effective Richardson constant from A^* to A^{**}, by as much as 50%. Figure 18 shows the calculated room-temperature values of A^{**} for metal-Si systems with an impurity concentration of 10^{16} cm^{-3}. We note that for electrons (n-type Si), A^{**} in the field range 10^4 to 2×10^5 V/cm remains essentially at a constant value of about 110 A/cm^2-K^2. For holes (p-type Si), A^{**} in this field range also remains essentially constant but at a considerably lower value ($\approx$ 30 A/cm^2-K^2). For n-type GaAs, A^{**} has been calculated to be 4.4 A/cm^2-K^2.

We conclude from the foregoing discussions that at room temperature in the electric field range of 10^4 to about 10^5 V/cm, the current transport mechanism in most Si and GaAs Schottky-barrier diodes is mainly due to thermionic emission of majority carriers. The spatial dependence of the electron Fermi level E_{Fn} near the metal-semiconductor interface has been studied by substituting Eqs. 6 and 74 into Eq. 73 and evaluating the difference, $E_{Fn}(W_D) - E_{Fn}(0)$. The E_{Fn} as shown in Fig. 16 is essentially flat throughout the depletion region.[38] The difference $E_{Fn}(W_D) - E_{Fn}(0)$ for a Au-Si diode with $N_D = 1.2\times10^{15}$ cm^{-3}, is only 8 meV for a forward bias of 0.2 V at 300 K. At higher doping levels the difference is even smaller. These results further confirm that for high-mobility semiconductors with moderate dopings, the thermionic-emission theory is applicable.

3.3.4 Tunneling Current

For more heavily doped semiconductors and/or for operation at low temperatures, the tunneling current may become more significant. In the extreme of an ohmic contact,

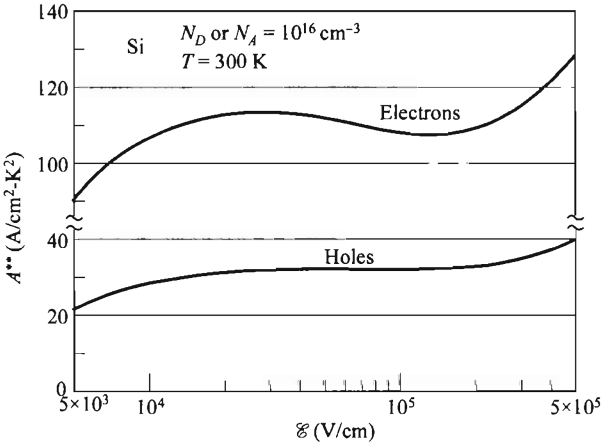


Fig. 18 Calculated effective Richardson constant A^{**} versus electric field for metal-silicon barriers. (After Ref. 37.)

which is a metal contact on degenerate semiconductor, the tunneling current is the dominant transport process. We will concentrate on ohmic contacts in the last section of this chapter.

The tunneling current from semiconductor to metal $J_{s\to m}$ is proportional to the quantum transmission coefficient (tunneling probability) multiplied by the occupation probability in the semiconductor and the unoccupied probability in the metal, that is,[36]

$$J_{s\to m} = \frac{A^{**}T^2}{kT}\int_{E_{Fm}}^{q\phi_{Bn}} F_s T(E)(1-F_m)dE. \tag{86}$$

F_s and F_m are the Fermi-Dirac distribution functions for the semiconductor and the metal respectively, and $T(E)$ is the tunneling probability which depends on the width of the barrier at a particular energy. A similar expression can be given for the current $J_{m\to s}$ which traverses in the opposite direction. In that case F_s and F_m would be interchanged in using the same equation. The net current density is the algebraic sum of the two components. Further analytical expression for the above equation is difficult, and the results can be obtained by numerical evaluation by computer.

Theoretical and experimental values of typical current-voltage characteristics for Au-Si barriers are shown in Fig. 19. We note that the total current density, which consists of both thermionic emission and tunneling, can be conveniently expressed as

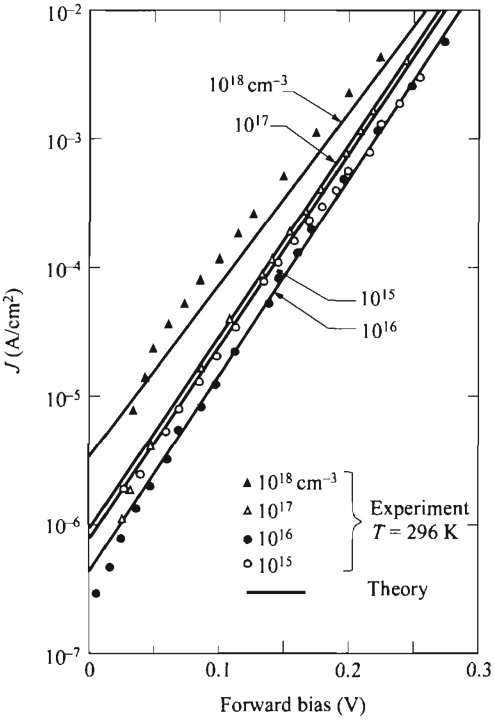


Fig. 19 Theoretical and experimental current-voltage characteristics for Au-Si Schottky barriers. Increased current is due to tunneling. (After Ref. 36.)

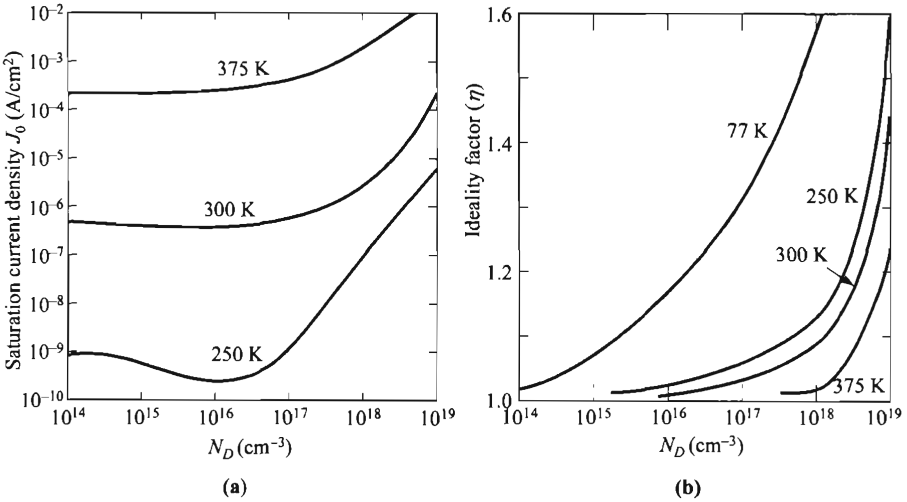
$$J = J_0\left[\exp\left(\frac{qV}{\eta kT}\right) - 1\right] \tag{87}$$

where J_0 is the saturation current density obtained by extrapolating the current density from the log-linear plot to $V = 0$, and η is the ideality factor, related to the slope. With little or no tunneling current or depletion-layer recombination, J_0 is determined by that of thermionic emission and η is close to unity. For higher doping and/or lower temperature, tunneling starts to occur and both J_0 and η increase.

The saturation current density J_0 and η are plotted in Fig. 20 for Au-Si diodes as a function of doping concentration, with temperature as a parameter. Note that J_0 is essentially a constant for low dopings but begins to increase rapidly when $N_D > 10^{17}$ cm^{-3}. The ideality factor η is very close to unity at low dopings and high temperatures. However, it can depart substantially from unity when the doping is increased or the temperature is lowered.

Figure 21 shows the ratio of the tunneling current to the thermionic current of a Au-Si barrier diode. Note that for $N_D \leq 10^{17}$ cm^{-3} and $T \geq 300$ K, the ratio is much less than unity and the tunneling component can be neglected. However, for higher dopings and lower temperatures, the ratio can become much larger than unity, indicating that the tunneling current becomes dominant.

Alternatively, the tunneling current can be expressed analytically and will give more physical insight. This formulation, based on the work of Padovani and Stratton,[39] is also used to derive the ohmic contact resistance. Referring to the energy-band diagrams in Fig. 22, we can roughly categorize the components into three types:



(a) (b)

Fig. 20 (a) Saturation current density versus doping concentration for Au-Si Schottky barriers at three temperatures. (b) Ideality factor η versus doping concentration at different temperatures. (After Ref. 36.)

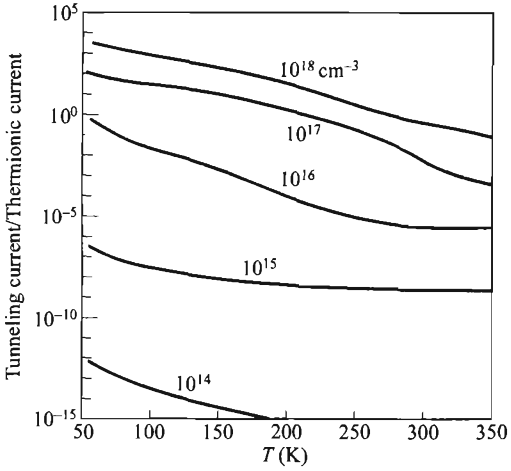


Fig. 21 Ratio of tunneling-current component to the thermionic-current component of a Au-Si barrier. The tunneling current will dominate at higher dopings and lower temperatures. (After Ref. 36.)

(1) thermionic emission (TE) over the barrier, (2) field emission (FE) near the Fermi level, and (3) thermionic-field emission (TFE) at an energy between TE and FE. While FE is a pure tunneling process, TFE is tunneling of thermally excited carriers which see a thinner barrier than FE. The relative contributions of these components depend on both temperature and doping level. A rough criterion can be set by comparing the thermal energy kT to E_{00} which is defined as

$$E_{00} \equiv \frac{q\hbar}{2}\sqrt{\frac{N}{m^* \varepsilon_s}}. \tag{88}$$

When $kT \gg E_{00}$, TE dominates and the original Schottky-barrier behavior prevails without tunneling. When $kT \ll E_{00}$, FE (or tunneling) dominates. When $kT \approx E_{00}$, TFE is the main mechanism which is a combination of TE and FE.

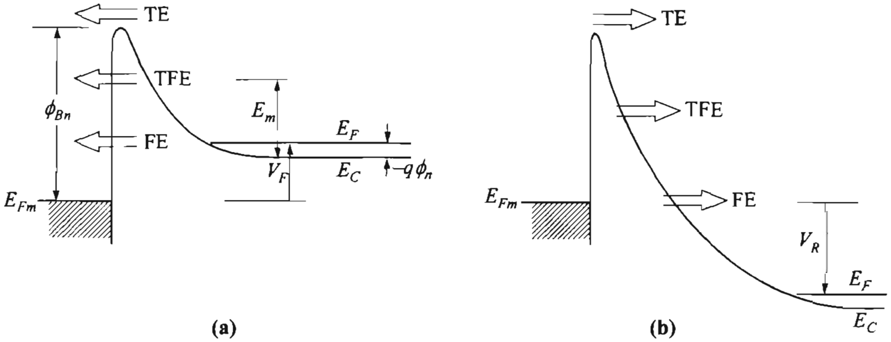


Fig. 22 Energy-band diagrams showing qualitatively tunneling currents in a Schottky diode (on *n*-type degenerate semiconductor) under (a) forward bias and (b) reverse bias. TE = thermionic emission. TFE = thermionic-field emission. FE = field emission.

Under forward bias, the current due to FE can be expressed as[39]

$$J_{FE} = \frac{A^{**}T\pi \exp[-q(\phi_{Bn}-V_F)/E_{00}]}{c_1 k \sin(\pi c_1 kT)}[1 - \exp(-c_1 qV_F)]$$
$$\approx \frac{A^{**}T\pi \exp[-q(\phi_{Bn}-V_F)/E_{00}]}{c_1 k \sin(\pi c_1 kT)} \tag{89}$$

where

$$c_1 \equiv \frac{1}{2E_{00}} \log\left[\frac{4(\phi_{Bn} - V_F)}{-\phi_n}\right]. \tag{90}$$

(ϕ_n is negative for degenerate semiconductors.) Notice the much weaker temperature dependence here (absent in the exponential term) compared to TE which is a characteristic of tunneling. The current due to TFE is given by

$$J_{TFE} = \frac{A^{**}T\sqrt{\pi E_{00} q(\phi_{Bn} - \phi_n - V_F)}}{k \cosh(E_{00}/kT)} \exp\left[\frac{-q\phi_n}{kT} - \frac{q(\phi_{Bn} - \phi_n)}{E_0}\right] \exp\left(\frac{qV_F}{E_0}\right), \tag{91}$$

$$E_0 \equiv E_{00} \coth\left(\frac{E_{00}}{kT}\right). \tag{92}$$

This TFE peaks roughly at an energy

$$E_m = \frac{q(\phi_{Bn} - \phi_n - V_F)}{\cosh^2(E_{00}/kT)} \tag{93}$$

where E_m is measured from E_C of the neutral region.

Under reverse bias, the tunneling current can be much larger because a large voltage is possible. The currents due to FE and TFE are given by

$$J_{FE} = A^{**}\left(\frac{E_{00}}{k}\right)^2 \left(\frac{\phi_{Bn} + V_R}{\phi_{Bn}}\right) \exp\left(-\frac{2q\phi_{Bn}^{3/2}}{3E_{00}\sqrt{\phi_{Bn} + V_R}}\right), \tag{94}$$

$$J_{TFE} = \frac{A^{**}T}{k} \sqrt{\pi E_{00} q\left[V_R + \frac{\phi_{Bn}}{\cosh^2(E_{00}/kT)}\right]} \exp\left(\frac{-q\phi_{Bn}}{E_0}\right) \exp\left(\frac{qV_R}{\varepsilon'}\right), \tag{95}$$

where

$$\varepsilon' = \frac{E_{00}}{(E_{00}/kT) - \tanh(E_{00}/kT)}. \tag{96}$$

These analytical expressions, although complicated, can be easily evaluated if all the parameters are known. These equations are also used to derive the ohmic contact resistance in the last section of this chapter.

3.3.5 Minority-Carrier Injection

The Schottky-barrier diode is mainly a majority-carrier device. The minority-carrier injection ratio γ, which is the ratio of minority-carrier current to total current, is small because the minority-carrier diffusion is much smaller than the majority-carrier thermionic-emission current. However, at sufficiently large forward bias, the drift component of the minority carriers cannot be ignored anymore and the increased drift component will increase the overall injection efficiency. Both drift and diffusion of holes lead to the total current of

$$J_p = q\mu_p p_n \mathscr{E} - qD_p\frac{dp_n}{dx}. \tag{97}$$

The increased field is set up by the large majority-carrier thermionic-emission current,

$$J_n = q\mu_n N_D \mathscr{E}. \tag{98}$$

We consider the energy-band diagram shown in Fig. 23 where x_1 is the boundary of the depletion layer, and x_2 marks the interface between the n-type epitaxial layer and the n^+-substrate. From the junction theory discussed in Chapter 2, the minority-carrier density at x_1 is

$$p_n(x_1) = p_{no}\exp\left(\frac{qV}{kT}\right) = \frac{n_i^2}{N_D}\exp\left(\frac{qV}{kT}\right). \tag{99}$$

This quantity $p_n(x_1)$ can also be expressed as a function of the forward current density, obtained from Eqs. 84 and 99:

$$p_n(x_1) \approx \frac{n_i^2}{N_D}\frac{J_n}{J_{n0}}, \tag{100}$$

where J_{n0} (saturation current density) and J_n are representations of the thermionic-emission current (Eq. 84) in the following form:

$$J_n = J_{n0}\exp\left[\left(\frac{qV}{kT}\right) - 1\right]. \tag{101}$$

The other boundary condition for $p_n(x_2)$ is also necessary to calculate the diffusion current. We use the term transport velocity S_p (or surface recombination velocity) for the minority carriers to relate the current and concentration by

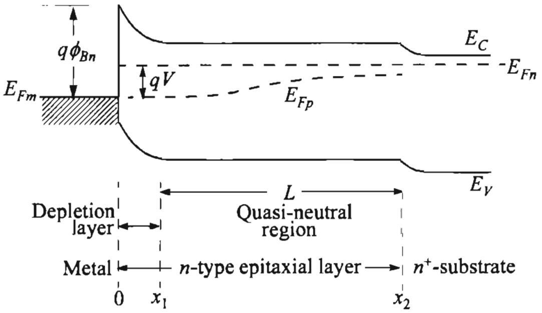


Fig. 23 Energy-band diagram of an epitaxial Schottky barrier under forward bias.

$$J_p(x_2) = qS_p[p_n(x_2) - p_{no}]. \tag{102}$$

We first consider the case with $S_p = \infty$ or equivalently $p_n(x_2) = p_{no}$. Under this boundary condition, the diffusion component has a standard form as in a *p-n* junction. From Eqs. 97, 98, and 100 we obtain the total hole current as (for $L \ll L_p$)

$$\begin{aligned} J_p &= q\mu_p p_n \mathscr{E} + \frac{qD_p n_i^2}{N_D L}\exp\left[\left(\frac{qV}{kT}\right) - 1\right] \\ &= \frac{\mu_p n_i^2 J_n^2}{\mu_n N_D^2 J_{n0}} + \frac{qD_p n_i^2}{N_D L}\exp\left[\left(\frac{qV}{kT}\right) - 1\right]. \end{aligned} \tag{103}$$

The injection ratio is given by

$$\gamma \equiv \frac{J_p}{J_p + J_n} \approx \frac{J_p}{J_n} \approx \frac{\mu_p n_i^2 J_n}{\mu_n N_D^2 J_{n0}} + \frac{qD_p n_i^2}{N_D L J_{n0}}. \tag{104}$$

For Au-Si diodes, the injection ratio has been measured to be very small, of the order of 10^{-5}, in agreement with the above equation.[40] Notice that γ has two terms. The second term is due to diffusion and is bias independent. This is the injection ratio for low-level bias,

$$\gamma_0 = \frac{qD_p n_i^2}{N_D L J_{n0}}. \tag{105}$$

The first term is due to the drift process, and is bias (or current) dependent. It can surpass the diffusion component at high current.

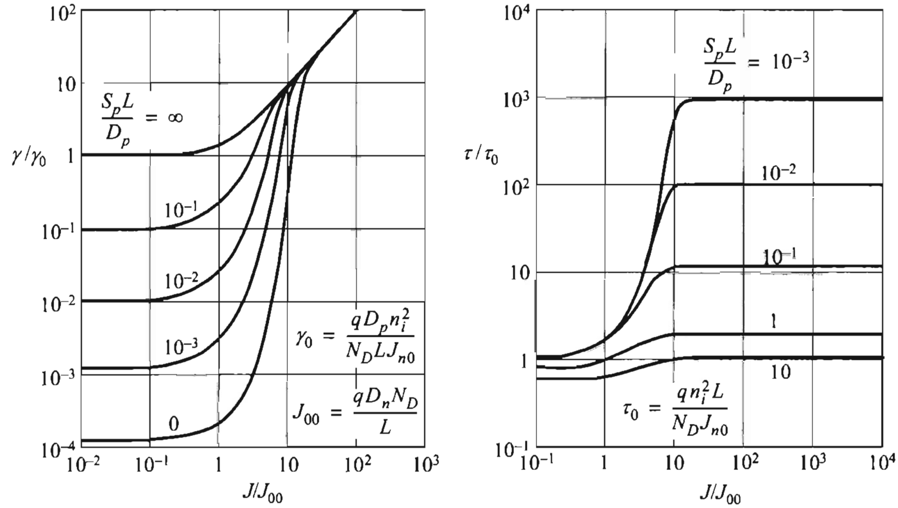
It is evident that to reduce the minority-carrier injection ratio (to reduce the charge storage time to be discussed below) one must use a metal-semiconductor system with large N_D (corresponding to low resistivity material), large J_{n0} (corresponding to small barrier height), and small n_i (corresponding to large bandgap). Furthermore, high-level bias is to be avoided. As an example, a gold-*n*-silicon diode with $N_D = 10^{15}$ cm^{-3} and $J_{n0} = 5\times10^{-7}$ A/cm^2 would give a low-bias injection γ_0 of $\approx 5\times10^{-4}$. But it would be expected to have an injection ratio of about 5% at a current density of 350 A/cm^2.

The above assumes that $p_n(x_2) = p_{no}$. Notice that at x_2, there is a barrier for holes that causes the holes to build up. These intermediate cases have been considered by Scharfetter using S_p as a parameter.[41] The computed results are shown in Fig. 24a, where the normalization factors are given by γ_0 and

$$J_{00} \equiv \frac{qD_n N_D}{L}. \tag{106}$$

J_{00} is the majority-carrier current at which the hole drift and diffusion components become equal, obtained by equating the two terms in Eq. 103.

Another quantity associated with the injection ratio is the minority-carrier storage time τ_s, which is defined as the minority carrier stored in the quasi-neutral region per unit current density:



(a) **(b)**

Fig. 24 (a) Normalized minority-carrier injection ratio versus normalized current density. (b) Normalized minority-carrier storage time versus normalized current density. $L/L_p = 10^{-2}$. (After Ref. 41.)

$$\tau_s \equiv \int_{x_1}^{x_2} qp(x)dx \Big/ J \tag{107}$$

For the low-current limit, depending on $p_n(x_2)$ or S_p, τ_s is given approximately by (for $L \ll L_p$)

$$\tau_s \approx \frac{qn_i^2L}{N_DJ_{n0}}, \tag{108}$$

and is independent of current. For high-current biases, $p_n(x_2)$ can become much higher, even to the extent that it is larger than in the rest of the quasi-neutral region L, i.e. a profile that increases with distance. The general results, using S_p again as a parameter, for τ_s versus the current density are shown in Fig. 24b. It can be seen that for finite S_p ($S_p \neq \infty$), τ_s can increase by orders of magnitude. Also, a high doping level is critical to reduce storage time in all cases.

3.3.6 MIS Tunnel Diode

In the *metal-insulator-semiconductor* (MIS) tunnel diode, a thin interfacial layer such as an oxide is intentionally (or sometimes unintentionally) introduced before metal deposition.[42,43] This interfacial-layer thickness lies in the range of 1–3 nm. This device differs from the MIS capacitor (to be considered in Chapter 4) in having appreciable current and under bias the semiconductor is not in equilibrium, i.e., the quasi Fermi levels for electrons E_{Fn} and holes E_{Fp} split. The major differences of this structure compared to a conventional metal-semiconductor contact are: (1) reduced current because of the added interfacial layer, (2) lower barrier height (some potential is developed across the interfacial layer), and (3) higher ideality factor η. The energy-band diagram is similar to Fig. 5.

The current equation can be written as[42]

$$J = A^* T^2 \exp(-\sqrt{\zeta}\,\delta) \exp\left(\frac{-q\phi_B}{kT}\right)\left[\exp\left(\frac{qV}{\eta kT}\right) - 1\right]. \tag{109}$$

The derivation for this equation can be found in Section 8.3.2. For the same barrier, the current is suppressed by the tunneling probability $\exp(-\sqrt{\zeta}\,\delta)$. Here ζ (in eV) and δ (in Å) are the effective barrier and thickness of the interfacial layer. (A constant of $[2(2m^*/\hbar^2)]^{1/2}$ which has the value of 1.01 $eV^{-1/2}Å^{-1}$ is omitted.) This added tunneling probability can be considered as a modification to the effective Richardson constant, as discussed before. The ideality factor is increased to[42]

$$\eta = 1 + \left(\frac{\delta}{\varepsilon_i}\right)\frac{(\varepsilon_s/W_D) + qD_{its}}{1 + (\delta/\varepsilon_i)qD_{itm}} \tag{110}$$

where D_{its} and D_{itm} are interface traps in equilibrium with the semiconductor and metal, respectively. In general, when the oxide thickness is less than 3 nm, the interface traps are in equilibrium with the metal, whereas for thicker oxides, these traps tend to be in equilibrium with the semiconductor.

The interfacial layer reduces the majority-carrier thermionic-emission current without affecting the minority-carrier current, which is from diffusion, and raises the minority injection efficiency. This phenomenon is exploited in improving the injection efficiency of an electroluminescent diode and the open-circuit voltage of the Schottky-barrier solar cell.

3.4 MEASUREMENT OF BARRIER HEIGHT

Basically, four methods are used to measure the barrier height of a metal-semiconductor contact: the (1) current-voltage, (2) activation-energy, (3) capacitance-voltage, and (4) photoelectric methods.

3.4.1 Current-Voltage Measurement

For moderately doped semiconductors, the I-V characteristics in the forward direction with $V > 3kT/q$ is given by Eq. 84:

$$J = A^{**} T^2 \exp\left(-\frac{q\phi_{B0}}{kT}\right) \exp\left[\frac{q(\Delta\phi + V)}{kT}\right]. \tag{111}$$

Since both A^{**} and $\Delta\phi$ (image-force lowering) are weak functions of the applied voltage, the forward J-V characteristic (for $V > 3kT/q$) is represented by $J = J_0 \exp(qV/\eta kT)$, as given previously in Eq. 87, where η is the ideality factor:

$$\eta \equiv \frac{q}{kT} \frac{dV}{d(\ln J)}$$
$$= \left[1 + \frac{d\Delta\phi}{dV} + \frac{kT}{q}\frac{d(\ln A^{**})}{dV}\right]^{-1}. \tag{112}$$

Typical examples are shown in Fig. 25, where $\eta = 1.02$ for the W-Si diode and $\eta = 1.04$ for the W-GaAs diode. The extrapolated value of current density at zero voltage is the saturation current J_0, and the barrier height can be obtained from the equation

$$\phi_{Bn} = \frac{kT}{q}\ln\left(\frac{A^{**}T^2}{J_0}\right). \tag{113}$$

The value of ϕ_{Bn} is not very sensitive to the choice of A^{**}, since at room temperature, a 100% increase in A^{**} will cause an increase of only 0.018 V in ϕ_{Bn}. The theoretical relationship between J_0 and ϕ_B (ϕ_{Bn} or ϕ_{Bp}) at room temperature is plotted in Fig. 26 for $A^{**} = 120$ A/cm^2-K^2. For other values of A^{**}, parallel lines can be drawn on this plot to obtain the proper relationship.

In the reverse direction, the dominant voltage dependence is due mainly to the Schottky-barrier lowering, or

$$J_R \approx J_0 \qquad \text{(for } V_R > 3kT/q\text{)}$$
$$\approx A^{**}T^2 \exp\left[-\frac{q(\phi_{B0} - \sqrt{q\mathscr{E}_m/4\pi\varepsilon_s})}{kT}\right] \tag{114}$$

where

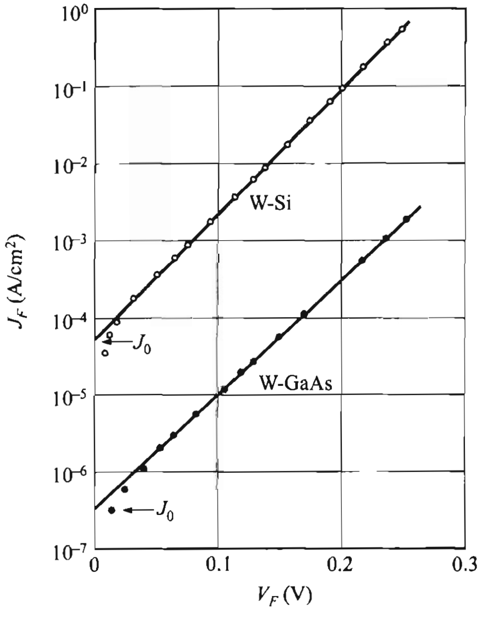


Fig. 25 Forward current density versus applied voltage of W-Si and W-GaAs diodes. (After Ref. 44.)

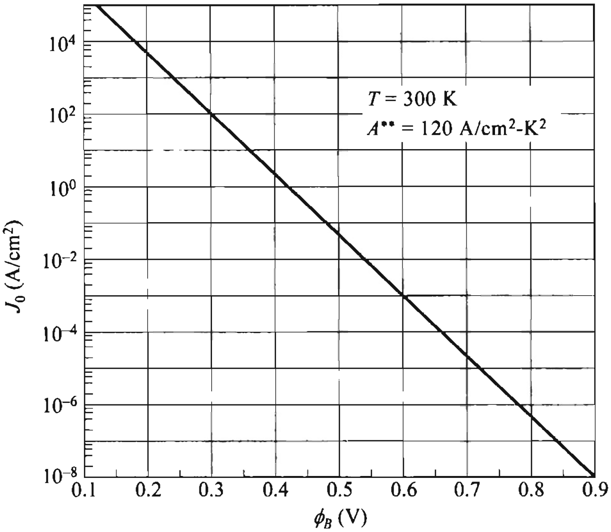


Fig. 26 Theoretical saturation current density at 300 K versus barrier height for an effective Richardson constant of 120 A/cm^2-K^2.

$$\mathscr{E}_m = \sqrt{\frac{2qN_D}{\varepsilon_s}\left(V_R + \psi_{bi} - \frac{kT}{q}\right)}. \tag{115}$$

If the barrier height $q\phi_{Bn}$ is sufficiently smaller than the bandgap so that the depletion-layer generation-recombination current is small in comparison with the Schottky emission current, then the reverse current will increase gradually with the reverse bias as given by Eq. 114, due mainly to image-force lowering.

For most of the practical Schottky diodes, however, the dominant reverse current component is the edge leakage current, which is caused by the sharp edge around the periphery of the metal plate. This sharp-edge effect is similar to the junction-curvature effect (with $r_j \rightarrow 0$) as discussed in Chapter 2. To eliminate this effect, metal-semiconductor diodes have been fabricated with a diffused guard ring (these structures will be discussed later). The guard ring is a deep p-type diffusion, and the doping profile is tailored to give the p-n junction a higher breakdown voltage than that of the metal-semiconductor contact. Because of the elimination of the sharp-edge effect, near-ideal forward and reverse I-V characteristics have been obtained. Figure 27 shows a comparison between experimental measurement from a PtSi-Si diode with guard ring, and theoretical calculation based on Eq. 114. The agreement is excellent. The sharp increase of current near 30 V is due to avalanche breakdown and is expected for the diode with a donor concentration of 2.5×10^{16} cm^{-3}.

The efficacy of guard ring structures in preventing premature breakdown and surface leakage can be ascertained by studying the reverse leakage current as a function of diode diameter at constant reverse bias. For this purpose, arrays of Schottky

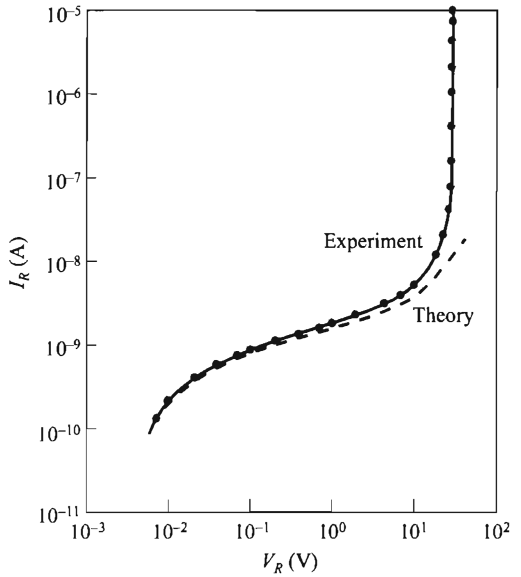


Fig. 27 Comparison of measurement with the theoretical prediction of reverse-bias current from Eq. 114 for a PtSi-Si diode. (After Ref. 45.)

diodes with different diameters can be formed on the semiconductor. The reverse leakage current can be measured and plotted as a function of diode diameter.[46] If the experimental data have a slope equal to two, the leakage currents are proportional to the device area. If, on the other hand, the leakage currents are dominated by edge effects, the data would be expected to lie along a straight line with a slope equal to unity.

For some Schottky diodes, the reverse current has an additional component. This component arises from the fact that if the metal-semiconductor interface is free from intervening layers of oxide and other contaminants, the electrons in the metal have wave functions that penetrate into the semiconductor energy gap. This is a quantum-mechanical effect that results in a static dipole layer at the metal-semiconductor interface. The dipole layer causes the intrinsic barrier height to vary slightly with the field, so $d\phi_{B0}/d\mathscr{E}_m \neq 0$. To a first approximation the static lowering can be expressed as

$$\Delta\phi_{\text{static}} \approx \alpha\mathscr{E}_m \tag{116}$$

or $\alpha \equiv d\phi_{B0}/d\mathscr{E}_m$. Figure 28 shows good agreement between the theory and measurements of the reverse current in a RhSi-Si diode, based on an empirical value of $\alpha = 1.7$ nm.

3.4.2 Activation-Energy Measurement

The principal advantage of Schottky-barrier determination by means of an activation energy measurement is that no assumption of electrically active area is required. This feature is particularly important in the investigation of novel or unusual metal-semiconductor interfaces because often the true value of the contacting area is not known. In the case of poorly cleaned or incompletely reacted surfaces, the electrically active

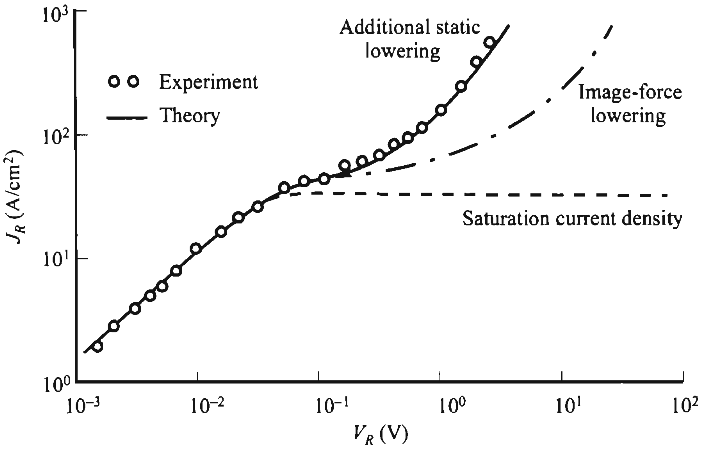


Fig. 28 Theory and experimental results of reverse characteristics for a RhSi-Si diode. (After Ref. 37.)

area may be only a small fraction of the geometric area. On the other hand, a strong metallurgical reaction could result in rough nonplanar metal-semiconductor interface with an electrically active area that is larger than the apparent geometric area.

If Eq. 84 is multiplied by A, the electrically active area, we obtain

$$\ln\left(\frac{I_F}{T^2}\right) = \ln(AA^{**}) - \frac{q(\phi_{Bn} - V_F)}{kT} \tag{117}$$

where $q(\phi_{Bn} - V_F)$ is considered the activation energy. Over a limited range of temperature around room temperature, the value of A^{**} and ϕ_{Bn} are essentially temperature independent. Thus for a fixed forward bias V_F, the slope of a plot of $\ln(I_F/T^2)$ versus $1/T$ yields the barrier height ϕ_{Bn}, and the ordinate intercept at $1/T = 0$ yields the product of the electrically active area A and the effective Richardson constant A^{**}.

To illustrate the importance of the activation-energy method in the investigation of interfacial metallurgical reactions, Fig. 29 shows the activation-energy plots of the saturation current in Al-n-Si contacts of different barrier heights, formed simply by annealing at various temperatures.[47] The slopes of these plots indicate a nearly linear increase of effective Schottky barrier height from 0.71 to 0.81 V for annealing temperatures between 450°C and 650°C. These observations were also confirmed with I-V and C-V measurements. Also supposedly when the Al-Si eutectic temperature ($\approx$ 580°C) is reached, the true metallurgical nature of the metal-semiconductor interface must be considerably modified. Determination of the ordinate intercepts from the plots shown in Fig. 29 indicates that the electrically active area increases by a factor of two, when the annealing temperature exceeds the Al-Si eutectic temperature.

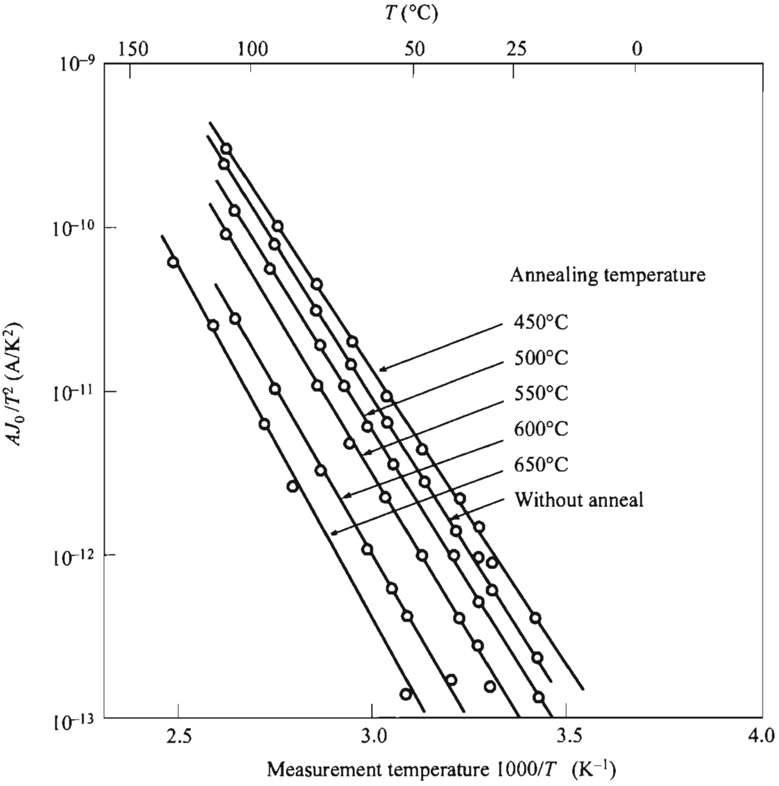


Fig. 29 Activation energy plots for determination of barrier height. (After Ref. 47.)

3.4.3 Capacitance-Voltage Measurement

The barrier height can also be determined by the capacitance measurement. When a small ac voltage is superimposed upon a dc bias, incremental charges of one sign are induced on the metal surface and charges of the opposite sign in the semiconductor. The relationship between C (depletion-layer capacitance per unit area) and V is given by Eq. 10. Figure 30 shows some typical results where $1/C^2$ is plotted against the applied voltage. The intercept on the voltage axis gives the built-in potential ψ_{bi} from which the barrier height can be determined:[44,48]

$$\phi_{Bn} = \psi_{bi} + \phi_n + \frac{kT}{q} - \Delta\phi. \tag{118}$$

From the slope the carrier density can also be determined (Eq. 11) and it can be used to calculate ϕ_n.

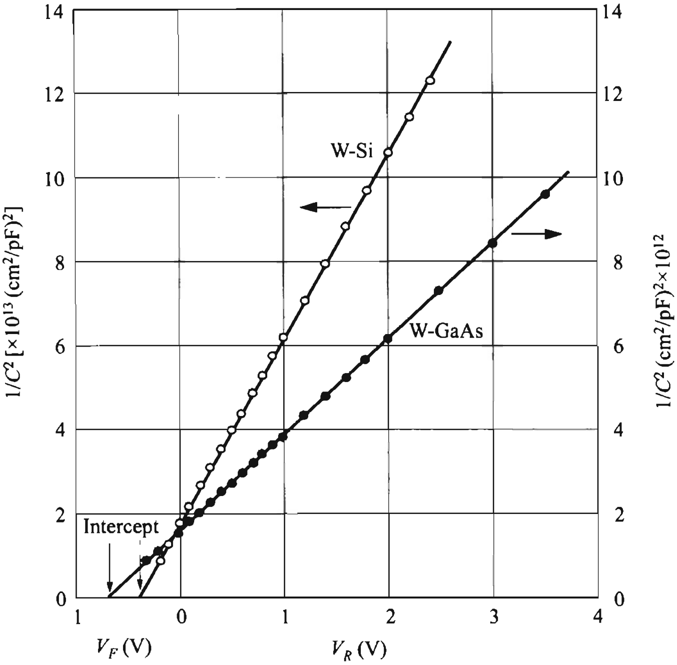


Fig. 30 $1/C^2$ versus applied voltage for W-Si and W-GaAs diodes. (After Ref. 44.)

To obtain the barrier height of semiconductor which contain both shallow-level and deep-level impurities (Fig. 4), we need to measure the C-V curves at two different temperatures at multiple frequencies.[49]

3.4.4 Photoelectric Measurement

The photoelectric measurement is an accurate and direct method of determining the barrier height.[50] When a monochromatic light is incident upon a metal surface, photocurrent may be generated. The basic setup is shown in Fig. 31. In a Schottky-barrier diode, two kinds of carrier excitation can occur that contribute to photocurrent; excitation over the barrier (process-1) and band-to-band excitation (process-2). In measuring the barrier height, only process-1 is useful and the most useful wavelengths should be in the range of $q\phi_{Bn} < h\nu < E_g$. Furthermore, the most critical light absorption region is at the metal-semiconductor interface. For front illumination, the metal film should be thin so light can penetrate to that interface. There is no such restriction in using back illumination since light is transparent in the semiconductor if $h\nu < E_g$, and the highest light intensity would be at the metal-semiconductor interface. Note that photocurrent can be collected without bias.

The photocurrent per absorbed photon (photoresponse R) as a function of the photon energy $h\nu$, is given by the Fowler theory:[51]

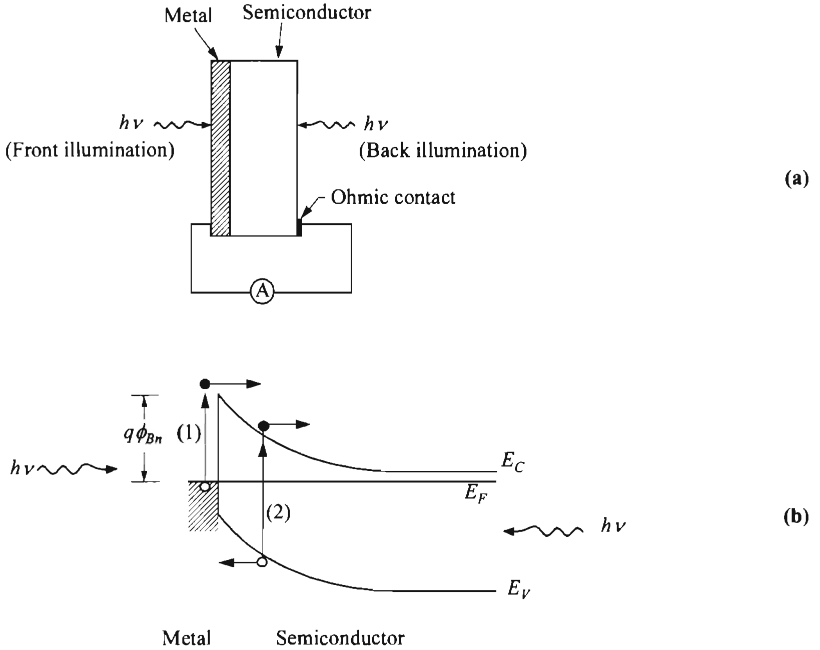

Metal
Semiconductor
$h\nu$
(Front illumination)
$h\nu$
(Back illumination)
Ohmic contact
A
(a)
$q\phi_{Bn}$
(1)
(2)
$h\nu$
E_C
E_F
$h\nu$
E_V
Metal
Semiconductor
(b)

Fig. 31 (a) Schematic setup for photoelectric measurement. (b) Energy-band diagram for photoexcitation processes.

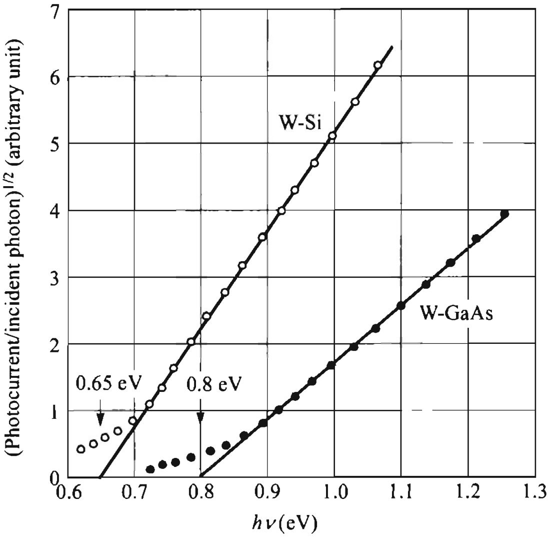


Fig. 32 Square root of the photoresponse versus photon energy for W-Si and W-GaAs diodes. The extrapolated values are the corresponding barrier height $q\phi_{Bn}$. (After Ref. 44.)

$$R \propto \frac{T^2}{\sqrt{E_s - h\nu}} \left\{ \frac{x^2}{2} + \frac{\pi^2}{6} - \left[\exp(-x) - \frac{\exp(-2x)}{4} + \frac{\exp(-3x)}{9} - \ldots \right] \right\} \quad \text{for } x \geq 0 \tag{119}$$

where E_s is the sum of $h\nu_0$ (= barrier height $q\phi_{Bn}$) and the Fermi energy measured from the bottom of the metal conduction band, and $x \equiv h(\nu - \nu_0)/kT$. Under the condition of $E_s \gg h\nu$ and $x > 3$, Eq. 119 reduces to

$$R \propto (h\nu - h\nu_0)^2 . \tag{120}$$

When the square root of the photoresponse is plotted as a function of photon energy, a straight line should be obtained, and the extrapolated value on the energy axis should give directly the barrier height. Figure 32 shows the photoresponse of W-Si and W-GaAs diodes, and the barrier heights of 0.65 and 0.80 eV are obtained respectively.

The photoelectric measurement can be used to study other device and material parameters. It has been used to determine the image-force dielectric constant of Au-Si diodes.[25] By measuring the shift of the photothreshold under different reverse biases, one can determine the image-force lowering $\Delta\phi$. From a plot of $\Delta\phi$ versus $\sqrt{\mathscr{E}_m}$, the dielectric constant ($\varepsilon_s/\varepsilon_0$) can be determined, as shown previously in Fig. 12. Photoelectric measurement has been used to study the temperature dependence of the barrier height.[52] The photothreshold is measured as a function of the temperature of Au-Si diodes. The shift of photothreshold correlates reasonably well with the temperature dependence of the silicon bandgap. This result implies that the Fermi level at the Au-Si interface is pinned in relation to the valence-band edge and this is in agreement with our discussion in Section 3.2.3.

3.4.5 Measured Barrier Heights

The *I-V*, *C-V*, activation-energy and photoelectric methods have been used to measure the barrier heights. For intimate contacts with a clean interface, these methods generally yield consistent barrier heights within ± 0.02 V. A large discrepancy between different methods may result from such causes as contamination in the interface, an intervening insulating layer, edge leakage current, or deep impurity levels.

The measured Schottky barrier heights for some elemental and compound semiconductors are listed in Table 3. The barrier heights are representative values for metal-semiconductor contacts made by deposition of high-purity metals in a good vacuum system onto cleaved or chemically cleaned semiconductor surfaces. As expected, silicon and GaAs metal-semiconductor contacts are most extensively studied. Among the metals, gold, aluminum, and platinum are most commonly used. The barrier heights of metal silicides on *n*-type silicon and some of their properties are listed in Table 4.

It should be pointed out that the barrier height is generally sensitive to pre-deposition surface preparation and post-deposition heat treatments.[63] Figure 33 shows the barrier heights on *n*-type Si and GaAs measured at room temperature after annealing at various temperatures. When an Al-Si diode is annealed above 450°C, the barrier height begins to increase,[47] presumably due to diffusion of Si in Al (also see Fig. 29).

Table 3 Measured Schottky-Barrier Heights ϕ_{Bn} (V) at 300 K on *n*-type Semiconductors. Each Entry Represents the Highest Value Reported for that System. Barrier Heights on *p*-type Can Be Estimated by $\phi_{Bp} + \phi_{Bn} \approx E_g/q$ (After Refs. 8, 53–59)

	Si	GaAs	Ge	AlAs	SiC	GaP	GaSb	InP	ZnS	ZnSe	ZnO	CdS	CdSe	CdTe	PbO
E_g	1.12	1.42	0.66	2.16	3.0	2.24	0.67	1.29	3.6	2.82	3.2	2.43	1.7	1.6	
Ag	0.83	1.03	0.54			1.2	0.45	0.54	1.81	1.21		0.56	0.43	0.8	0.95
Al	0.81	0.93	0.48		1.3	1.06	0.6	0.5	0.8	0.75	0.68			0.76	
Au	0.83	1.05	0.59	1.2	1.4	1.3	0.61	0.52	2.2	1.51	0.65	0.78	0.7	0.86	
Bi		0.9					0.2			1.14				0.78	
Ca	0.4	0.56													
Co	0.81	0.86	0.5		1.4										
Cr	0.60	0.82			1.2	1.18		0.45							
Cu	0.8	1.08	0.5		1.3	1.2	0.47	0.42	1.75	1.1	0.45	0.5	0.33	0.82	
Fe	0.98	0.84	0.42							1.11				0.78	
Hf	0.58	0.82				1.84									
In		0.83	0.64				0.6		1.5	0.91	0.3			0.69	0.93
Ir	0.77	0.91	0.42												
Mg	0.6	0.66				1.04	0.3		0.82	0.49					
Mo	0.69	1.04			1.3	1.13									
Ni	0.74	0.91	0.49		1.4	1.27		0.32				0.45		0.83	0.96
Os	0.7		0.4									0.53			
Pb	0.79	0.91	0.38							1.15		0.59		0.68	0.95
Pd	0.8	0.93			1.2		0.6	0.41	1.87		0.68	0.62		0.86	
Pt	0.9	0.98		1.0	1.7	1.45			1.84	1.4	0.75	1.1	0.37	0.89	
Rh	0.72	0.90	0.4												
Ru	0.76	0.87	0.38												
Sb		0.86					0.42			1.34				0.76	
Sn		0.82						0.35							
Ta		0.85							1.1		0.3				
Ti	0.6	0.84			1.1	1.12						0.84			
W	0.66	0.8	0.48												

Table 4 Barrier Height of Metal Silicide on *n*-type Si. For Each System, the Barrier-Height Entry Represents the Highest Reported Value (After Refs. 8, 22, 56, 60–62)

Metal Silicide	ϕ_{Bn} (V)	Structure	Forming Temperature (°C)	Melting Temperature (°C)
CoSi	0.68	Cubic	400	1460
$CoSi_2$	0.64	Cubic	450	1326
$CrSi_2$	0.57	Hexagonal	450	1475
$DySi_2$	0.37			
$ErSi_2$	0.39			
$GdSi_2$	0.37			
HfSi	0.53	Orthorhombic	550	2200
$HoSi_2$	0.37			
IrSi	0.93		300	
Ir_2Si_3	0.85			
$IrSi_3$	0.94			
MnSi	0.76	Cubic	400	1275
$Mn_{11}Si_{19}$	0.72	Tetragonal	800[a]	1145
$MoSi_2$	0.69	Tetragonal	1000[a]	1980
Ni_2Si	0.75	Orthorhombic	200	1318
NiSi	0.75	Orthorhombic	400	992
$NiSi_2$	0.66	Cubic	800[a]	993
Pd_2Si	0.75	Hexagonal	200	1330
PtSi	0.87	Orthorhombic	300	1229
Pt_2Si	0.78			
RhSi	0.74	Cubic	300	
$TaSi_2$	0.59	Hexagonal	750[a]	2200
$TiSi_2$	0.60	Orthorhombic	650	1540
VSi_2	0.65			
WSi_2	0.86	Tetragonal	650	2150
YSi_2	0.39			
$ZrSi_2$	0.55	Orthorhombic	600	1520

[a] Can be ≤ 700°C under clean interface condition.

Also, for metals that form silicides on silicon, the barrier height changes abruptly when the eutectic temperature is reached. The barrier height of a Pt-Si diode is 0.9 V. After annealing at 300°C or higher temperatures, PtSi is formed at the interface and ϕ_{Bn} decreases to 0.85 V.[64] For Pt-GaAs contact the barrier height increases from 0.84 V to 0.87 V when $PtAs_2$ is formed at the interface.[65] For a W-Si diode the barrier height remains constant until the annealing temperature is above 1000°C, when WSi_2 is formed.[66]

So far in all the Schottky diodes discussed above, the metal layers are deposited so they are polycrystalline or amorphous in structure. For certain silicide contacts on silicon, it has been demonstrated that single-crystalline form can be grown epitaxially

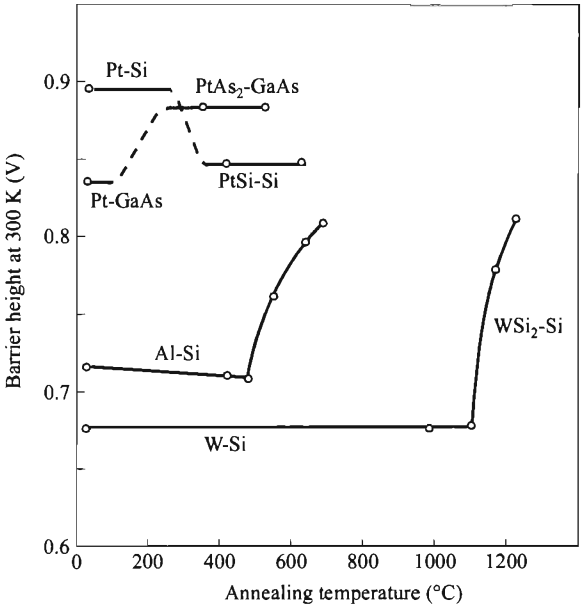


Fig. 33 Barrier heights on *n*-type Si and GaAs measured at room temperature after annealing at various temperatures.

from the underlying single-crystalline silicon.[67] These epitaxial silicides include $NiSi_2$, $CoSi_2$, $CrSi_2$, Pd_2Si, $ErSi_{2-x}$, $TbSi_{2-x}$, YSi_{2-x}, and $FeSi_2$. The epitaxial silicides have the properties of high uniformity and thermal stability. They provide a unique opportunity to study the fundamental relationship of barrier height to the microscopic interfacial configuration. It has been demonstrated that even on the same orientation of Si surface, different types (A and B) and interface structures (6-, 7-, or 8-folded) can be formed and that they give a difference in barrier height of as much as 0.14 eV. With this insight, the range of barrier heights observed on the same metal-semiconductor system can be rationalized, due to the statistical spacial distribution of the interfacial structure.

3.5 DEVICE STRUCTURES

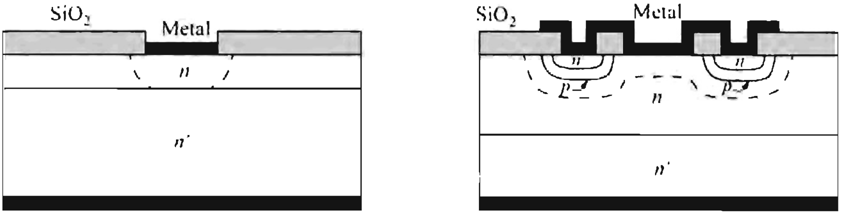
The earliest device structure is the point-contact rectifier using a small metal wire with a sharp point making contact with a semiconductor. The contact may be just a simple mechanical contact or it may be formed by an electrical discharge processes that may result in a small alloyed *p-n* junction.

A point-contact rectifier usually has poor forward and reverse *I-V* characteristics compared to a planar Schottky diode. Its characteristics are also difficult to predict from theory, since the rectifiers are subject to wide variations such as the whisker pressure, contact area, crystal structure, surface condition, whisker composition, and heat or forming processes. The advantage of a point-contact rectifier is its small area, which can give very small capacitance, a desirable feature for microwave application. The disadvantages are its large spreading resistance ($R_s \approx \rho/2\pi r_0$, where r_0 is the radius of the hemispheric point contact); large leakage current mainly due to the surface effect, which gives rise to poor rectification ratio, and soft reverse-breakdown characteristics due to a large concentrated field beneath the metal point.

Most modern metal-semiconductor diodes are made by a planar process. The metal-semiconductor contacts are formed by various methods including thermal evaporation (resistive or electron-beam heating), sputtering, chemical decomposition, or plating of metals. Surface preparation methods include chemical etch, mechanical polish, vacuum cleaving, back sputtering, heat treatment, or ion bombardment. Since most metal-semiconductor contacts are formed in a vacuum system,[68] an important parameter concerning vacuum deposition of metals is the vapor pressure, which is defined as the pressure exerted when a solid or liquid is in equilibrium with its own vapor.[69] Metals with high vapor pressure can be problematic during evaporation.

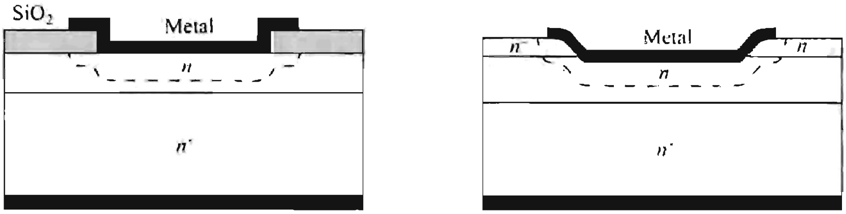
The most-common structures in integrated circuits have oxide isolation at the metal perimeter. The small-area contact device, Fig. 34a, fabricated by a planar process on epitaxial n on n^+-substrate, is useful as a microwave mixer diode.[70,71] To achieve good performance, we have to minimize the series resistance and the diode capacitance. The metal overlap structure,[72] Fig. 34b, gives near-ideal forward I-V characteristics and low leakage current at moderate reverse bias but the electrode sharp-edge effect will increase the reverse current when a large reverse bias is applied. This structure is extensively used in integrated circuits since it can be formed as an integral part of the metallization. Another approach is to use local-oxide isolation[73] to reduce the edge field, Fig. 34c. This approach requires a special planar process to incorporate a local oxidation step. In Fig. 34d, the diode is surrounded by a void or moat.[74] In this case, reliability problems can result from burying contaminants in the moat.

To eliminate the electrode sharp-edge effects, many device structures have been proposed. Figure 34e uses a diffused guard ring[45] to give near-ideal forward and reverse characteristics. This structure is useful as a tool to study static characteristics; however, it suffers from long recovery time and large parasitic capacitance due to the adjacent p-n junction. Figure 34f uses a double-diffused guard ring[75] to reduce the recovery time but the process is relatively complicated. Another guard-ring structure with a high resistivity layer on top of the active layer[76] is shown in Fig. 34g. Since the dielectric constant of the semiconductor is higher than that of an insulator, the parasitic capacitance is generally higher than the structure shown in Fig. 34b. The metal-overlap laterally diffused structure[77] is basically a double-Schottky diode (in parallel) that does not involve a p-n junction, Fig. 34h. This structure gives nearly ideal forward and reverse I-V characteristics with very short reverse recovery time. How-


SiO_2
Metal
n
n^+
SiO_2
Metal
n
p
p
n
n^+

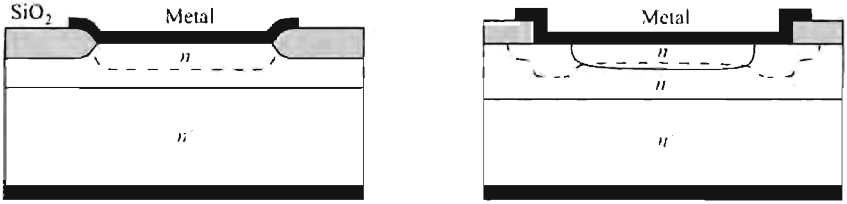
(a) Small-area contact

(f) Double-diffused guard ring


SiO_2
Metal
n
n^+
n^-
Metal
n^-
n
n^+

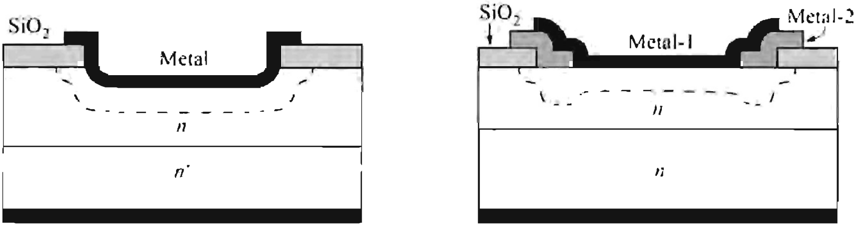
(b) Metal overlap

(g) High-resistivity guard layer


SiO_2
Metal
n
n^+
Metal
n
n
n^+

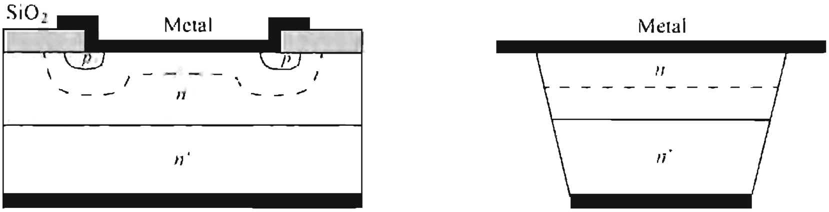
(c) Oxide isolated

(h) Metal-overlap laterally diffused


SiO_2
Metal
n
n^+
SiO_2
Metal-2
Metal-1
n
n

(d) Moat-etched

(i) Metal guard ring


SiO_2
Metal
p
p
n
n^+
Metal
n
n^+

(e) Diffused guard ring

(j) Truncated cone

Fig. 34 Various metal-semiconductor device structures. Depletion widths are indicated by dashed lines.

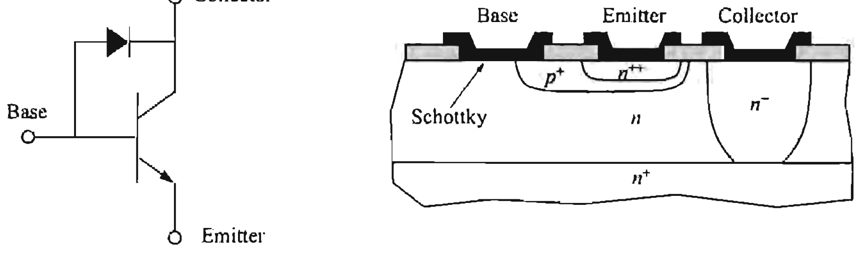
ever, the process involves extra oxidation and diffusion steps, and the outer n^--ring may increase the device capacitance.

The guard-ring structure, Fig. 34i, has been proposed that uses an additional metal with higher barrier height. However, large variations in barrier height are generally difficult to obtain for covalent semiconductors. For certain microwave power generators (e.g., IMPATT diodes) one uses the truncated-cone structure,[78] Fig. 34j. The angle between the metal overhang and the semiconductor cone must be larger than 90° so that the electric field at the contact periphery is always smaller than that in the center. This angle ensures that the avalanche breakdown will occur uniformly inside the metal-semiconductor contact.

One important application of Schottky diodes is the clamped bipolar transistor[79] (Fig. 35). (For a detailed discussion of the bipolar transistor, see Chapter 5.) A Schottky diode can be incorporated into the base-collector terminal to form a clamped (composite) transistor with a very short saturation time constant (see Section 5.3.3). Fabrication is simply achieved by allowing the base contact to extend to straddle the surrounding collector region in the standard buried-collector technology.[54] In the saturation region, the base-collector junction of the original transistor is slightly forward-biased instead of reverse-biased. If the forward voltage drop in the Schottky diode is much lower than the base-collector on-voltage of the original transistor, most of the excess base current flows through the Schottky diode which does not store minority carriers. Thus, the saturation time is reduced markedly compared with that of the original transistor.

Since a Schottky diode in general carries a larger current compared to other diodes, the series resistance is critical to this device. To characterize the series resistance, we start with a modified current equation from Eq. 87,

$$I = AJ_0\left\{\exp\left[\frac{q(V-IR_s)}{\eta kT}\right]-1\right\}. \tag{121}$$



(a) (b)

Fig. 35 Composite bipolar transistor (*npn*) with a Schottky diode clamp connected between the base and the *n*-collector. (a) Circuit representation. (b) Cross-section of structure.

From this, the differential resistance in the forward-bias regime is dependent on the bias or current, given by

$$\frac{dV}{dI} = \frac{\eta kT + qIR_s}{qI}. \tag{122}$$

This equation shows that the differential resistance of the diode at low bias is inversely proportional to the current ($= \eta kT/qI$). At high current when $IR_s \gg \eta kT/q$, the differential resistance would saturate to the value of R_s. Typical experimental results of differential resistance versus current are shown in Fig. 36a for Au-Si and Au-GaAs diodes. Also shown is the result for Si point contact, discussed previously. We note that for a sufficiently high forward bias the junction resistance approaches a constant value. This value is the series resistance R_s given by

$$R_s = \frac{1}{A}\int \rho(x)dx + \frac{\rho_S}{2\pi r}\tan^{-1}\left(\frac{2h}{r}\right) + R_{co}, \tag{123}$$

where the first term on the right is the resistance integrated over the quasi-neutral region (between the depletion-layer edge and the heavily doped substrate, as in Fig. 23). The second term is the spreading resistance in the substrate of resistivity ρ_S and thickness h, and a diode circular area of radius r (see last section). The last term R_{co} is the resistance due to the ohmic contact with the substrate. For a Schottky diode on a bulk semiconductor substrate, the first term is absent.

Another simple method to extract the series resistance is from the semilog plot of the I-V curve shown in Fig. 36b. In the region where the current deviates from the exponential rise, the resistance is estimated by $\Delta V = IR_s$.

An important figure of merit for microwave application of the Schottky diodes is the forward-bias cutoff frequency f_{c0}, which is defined as

$$f_{c0} \equiv \frac{1}{2\pi R_F C_F} \tag{124}$$

where R_F and C_F are the differential resistance and capacitance in a forward-bias range of ≈ 0.1 V to the flat-band condition.[81] The value of f_{c0} is considerably smaller than the corresponding cutoff frequency at zero bias, and can be used as a lower limit for practical consideration. A typical result is shown in Fig. 37. Note the higher cutoff frequency with a smaller junction diameter. Also for the same doping and junction diameter (e.g., 10 μm), the Schottky diode on n-type GaAs gives the highest cutoff frequency, mainly because the electron mobility is considerably higher, resulting in lower series resistance.

To improve the high-frequency performance, devices that have smaller capacitance but larger contact areas are desirable. It has been shown that the Mott barrier can meet this requirement. A Mott barrier is a metal-semiconductor contact in which the epitaxial layer is very lightly doped so that the whole epitaxial layer is fully depleted, resulting in low capacitance. This holds true even under forward bias so the capacitance remains constant, independent of bias. Figure 38 shows the band diagram of a Mott barrier. Since for a given cutoff frequency the capacitance is much lower than that of a standard Schottky diode, the Mott diode diameter can be made much

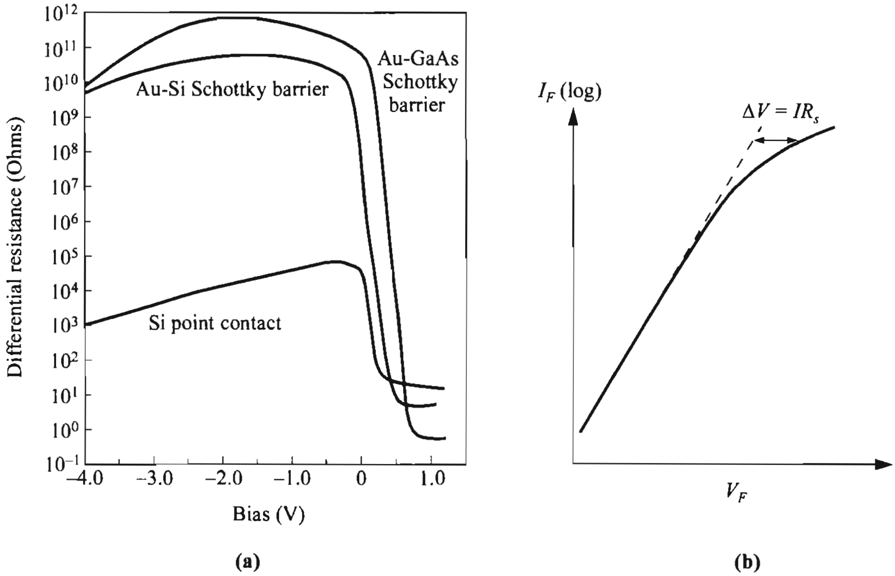


Fig. 36 (a) Measured differential resistance as a function of applied voltage for Au-Si, Au-GaAs, and point-contact diodes. (After Ref. 80.) (b) Estimate of series resistance from forward *I*-*V* curve.

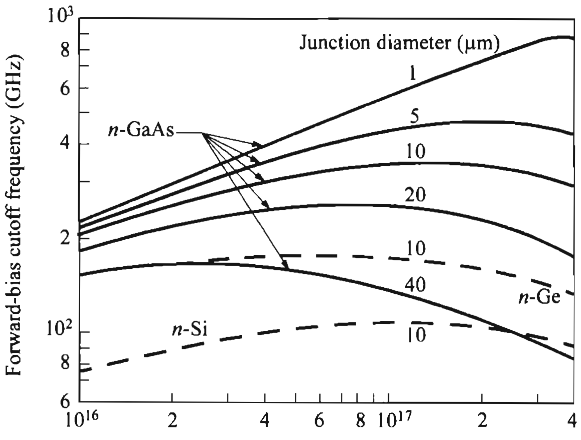


Fig. 37 Forward-bias cutoff frequency versus doping concentration in the epitaxial layer (0.5 μm thick) and with various junction diameters. (After Ref. 80.)

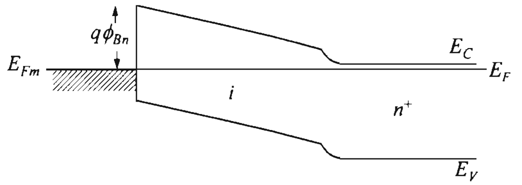


Fig. 38 Band diagram for Mott barrier at zero bias.

larger.[82] The current transport in a Mott barrier is dominated by diffusion, given by Eq. 72, due to the low majority-carrier concentration in the depletion region.

3.6 OHMIC CONTACT

An ohmic contact is defined as a metal-semiconductor contact that has a negligible junction resistance relative to the total resistance of the semiconductor device. A satisfactory ohmic contact should not significantly perturb the device performance and can supply the required current with a voltage drop that is sufficiently small compared with the drop across the active region of the device. The last connection to any semiconductor device is always an on-chip metallic layer. Thus, for every semiconductor device there are at least two metal-semiconductor contacts to form connections. So a good ohmic is a must for every semiconductor device.

The macroscopic parameter—specific contact resistance is defined as the reciprocal of the derivative of the current density with respect to the voltage across the interface. When evaluated at zero bias, this specific contact resistance R_c is an important figure-of-merit for ohmic contacts:[83]

$$R_c \equiv \left(\frac{dJ}{dV}\right)^{-1}_{V=0} . \tag{125}$$

Computer numerical simulation can be performed to get the solution.[83,84] Alternatively, to derive this R_c analytically, the I-V relationships described earlier in the chapter can be used. Again we use the comparison of doping (E_{00}) to temperature (kT) to decide which current mechanism is the dominant one.

For low to moderate doping levels and/or moderately high temperatures, $kT \gg E_{00}$, the standard thermionic-emission expression (Eq. 84) is used to obtain

$$R_c = \frac{k}{A^{**}Tq}\exp\left(\frac{q\phi_{Bn}}{kT}\right) \propto \exp\left(\frac{q\phi_{Bn}}{kT}\right) . \tag{126}$$

Since only small applied voltage is relevant, the voltage dependence of the barrier height can be neglected. Equation 126 shows that low barrier height should be used to obtain small R_c.

For higher doping level, $kT \approx E_{oo}$, TFE dominates and R_c is given by[39,85]

$$
\begin{aligned}
R_c &= \frac{k\sqrt{E_{00}}\cosh(E_{00}/kT)\coth(E_{00}/kT)}{A^{**}Tq\sqrt{\pi q(\phi_{Bn}-\phi_n)}}\exp\left[\frac{q(\phi_{Bn}-\phi_n)}{E_{00}\coth(E_{00}/kT)}+\frac{q\phi_n}{kT}\right] \\
&\propto \exp\left[\frac{q\phi_{Bn}}{E_{00}\coth(E_{00}/kT)}\right] . \qquad (127)
\end{aligned}
$$

(ϕ_n is negative for degenerate semiconductor.) This type of tunneling occurs at an energy above the conduction band where the product of carrier density and tunneling probability is at a maximum, given by E_m of Eq. 93.

With even higher doping level, $kT \ll E_{oo}$, FE dominates, and the specific contact resistance is given by[39,85]

$$
R_c = \frac{k\sin(\pi c_1 kT)}{A^{**}\pi q T}\exp\left(\frac{q\phi_{Bn}}{E_{00}}\right) \propto \exp\left(\frac{q\phi_{Bn}}{E_{00}}\right) . \qquad (128)
$$

Provided that the barrier height cannot be made very small, a good ohmic contact should operate in this regime of tunneling.

Specific contact resistance is a function of the barrier height (in all regimes), doping concentration (in TFE and FE), and temperature (more sensitive in TE and TFE). Qualitative dependence on these parameters is shown in Fig. 39 for a fixed semiconductor material. The trend and the regimes of operation are also indicated in the figure. In TE, R_c is independent of doping concentration and dependent only on the barrier height ϕ_B. In the other extreme of FE, in addition to ϕ_B, R_c has a dependence of $\propto \exp(N^{-1/2})$. The results of calculated specific contact resistance on silicon are given in Fig. 40.

It is quite obvious that to obtain low values of R_c, high doping concentration, low barrier height, or both must be used. And these are exactly the approaches used for all

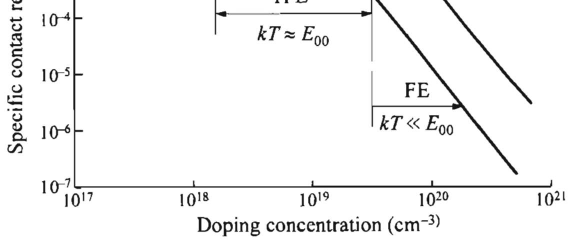
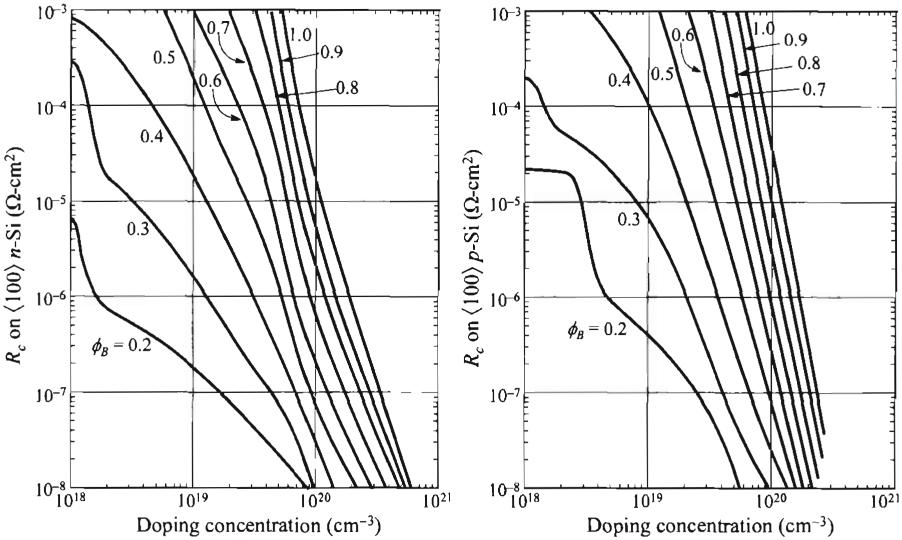


Fig. 39 Dependence of specific contact resistance on doping concentration (and E_{00}), barrier height, and temperature. Regimes of TE, TFE, and FE are indicated.



(a) **(b)**

Fig. 40 Calculated specific contact resistance R_c on (a) n-type and (b) p-type ⟨100⟩ Si surfaces for various barrier heights (in eV) at room temperature. (After Ref. 86)

ohmic contacts. On wide-gap semiconductors it is difficult to make good ohmic contacts. A metal does not generally exist with a low enough work function to yield a low barrier. In such cases, the general technique for making an ohmic contact involves the establishment of a more heavily doped surface layer. Another common technique is to add a heterojunction with a layer of small bandgap material and with high-level doping of the same type. For GaAs and other III-V compound semiconductors, various technologies have been developed for the ohmic contacts.[87] A summary of contact materials on common semiconductors is listed in Table 5.

As devices are miniaturized for advanced integrated circuits, the device current density usually increases. This demands not only smaller ohmic resistance but also a smaller contact area. The challenge of fabricating good ohmic contacts has been increasing with device miniaturization. The total contact resistance is given by

$$R = \frac{R_c}{A}. \tag{129}$$

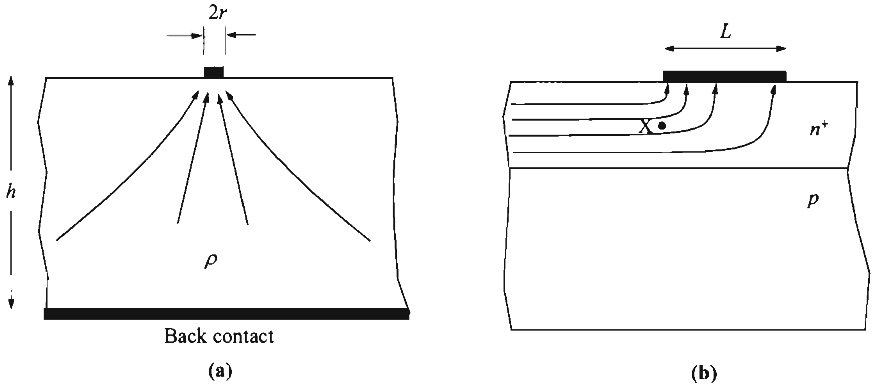
However, this expression is valid only for uniform current density across the whole area. We mention here two practical conditions that additional resistance components are important. For a small contact of radius r as shown in Fig. 41a, there is a spreading resistance in series with the ohmic contact given by[89]

$$R_{sp} = \frac{\rho}{2\pi r}\tan^{-1}\left(\frac{2h}{r}\right). \tag{130}$$

Table 5 Metal Ohmic Contacts for Various Semiconductors (After Ref. 88)

Semiconductor	Metal	Semiconductor	Metal
n-Ge	Ag-Al-Sb, Al, Al-Au-P, Au, Bi, Sb, Sn, Pb-Sn	*p*-Ge	Ag, Al, Au, Cu, Ga, Ga-In, In, Al-Pd, Ni, Pt, Sn
n-Si	Ag, Al, Al-Au, Ni, Sn, In, Ge-Sn, Sb, Au-Sb, Ti, TiN	*p*-Si	Ag, Al, Al-Au, Au, Ni, Pt, Sn, In, Pb, Ga, Ge, Ti, TiN
n-GaAs	Au(.88)Ge(.12)-Ni, Ag-Sn, Ag(.95)In(.05)-Ge	*p*-GaAs	Au(.84)Zn(.16), Ag-In-Zn, Ag-Zn
n-GaP	Ag-Te-Ni, Al, Au-Si, Au-Sn, In-Sn	*p*-GaP	Au-In, Au-Zn, Ga, In-Zn, Zn, Ag-Zn
n-GaAsP	Au-Sn	*p*-GaAsP	Au-Zn
n-GaAlAs	Au-Ge-Ni	*p*-GaAlAs	Au-Zn
n-InAs	Au-Ge, Au-Sn-Ni, Sn	*p*-InAs	Al
n-InGaAs	Au-Ge, Ni	*p*-InGaAs	Au-Zn, Ni
n-InP	Au-Ge, In, Ni, Sn		
n-InSb	Au-Sn, Au-In, Ni, Sn	*p*-InSb	Au-Ge
n-CdS	Ag, Al, Au, Au-In, Ga, In, Ga-In		
n-CdTe	In	*p*-CdTe	Au, In-Ni, Indalloy 13, Pt, Rh
n-ZnSe	In, In-Ga, Pt, InHg		
n-SiC	W	*p*-SiC	Al-Si, Si, Ni

This component approaches the bulk resistance of $\rho h/A$ for large r/h ratios. In cases where the contact is made on a horizontal diffusion layer (Fig. 41b, as in the case of a MOSFET), the total resistance between point X (leading edge of the contact) and the metal contact is given by[90]



(a) (b)

Fig. 41 (a) Current pattern for a small contact when $r \ll h$. r is the radius of the contact. (b) Current pattern for a contact to a horizontal diffusion sheet. If the sheet resistance of the diffusion layer is high, current is forced toward the leading edge of the contact.

$$R = \frac{\sqrt{R_{\square}R_c}}{W}\coth\left(L\sqrt{\frac{R_{\square}}{R_c}}\right) \tag{131}$$

where $R_{\square}$ is the sheet resistance ($\Omega/\square$) of the diffusion layer. This equation takes into account nonuniform current density through the contact (current crowding) and contributions due to the sheet resistance itself. It can also be shown that in the limit of $R_{\square} \to 0$, Eq. 131 reduces to Eq. 129.

REFERENCES

1. F. Braun, "Über die Stromleitung durch Schwefelmetalle," *Ann. Phys. Chem.*, **153**, 556 (1874).
2. J. C. Bose, U.S. Patent 775,840 (1904).
3. A. H. Wilson, "The Theory of Electronic Semiconductors," *Proc. R. Soc. Lond. Ser. A*, **133**, 458 (1931).
4. W. Schottky, "Halbleitertheorie der Sperrschicht," *Naturwissenschaften*, **26**, 843 (1938).
5. N. F. Mott, "Note on the Contact between a Metal and an Insulator or Semiconductor," *Proc. Cambr. Philos. Soc.*, **34**, 568 (1938).
6. H. A. Bethe, "Theory of the Boundary Layer of Crystal Rectifiers," *MIT Radiat. Lab. Rep.*, 43-12 (1942).
7. H. K. Henisch, *Rectifying Semiconductor Contacts*, Clarendon, Oxford, 1957.
8. E. H. Rhoderick and R. H. Williams, *Metal-Semiconductor Contacts*, 2nd Ed., Clarendon, Oxford, 1988.
9. E. H. Rhoderick, "Transport Processes in Schottky Diodes," in K. M. Pepper, Ed, *Inst. Phys. Conf. Ser.*, No. 22, Institute of Physics, Manchester, England, 1974, p. 3.
10. V. L. Rideout, "A Review of the Theory, Technology and Applications of Metal-Semiconductor Rectifiers," *Thin Solid Films*, **48**, 261 (1978).
11. R. T. Tung, "Recent Advances in Schottky Barrier Concepts," *Mater. Sci. Eng. R.*, **35**, 1 (2001).
12. H. B. Michaelson, "Relation between an Atomic Electronegativity Scale and the Work Function," *IBM J. Res. Dev.*, **22**, 72 (1978).
13. G. I. Roberts and C. R. Crowell, "Capacitive Effects of Au and Cu Impurity Levels in Pt *n*-type Si Schottky Barriers," *Solid-State Electron.*, **16**, 29 (1973).
14. A. M. Cowley and S. M. Sze, "Surface States and Barrier Height of Metal-Semiconductor Systems," *J. Appl. Phys.*, **36**, 3212 (1965).
15. J. Bardeen, "Surface States and Rectification at a Metal Semiconductor Contact," *Phys. Rev.*, **71**, 717 (1947).
16. C. A. Mead and W. G. Spitzer, "Fermi-Level Position at Metal-Semiconductor Interfaces," *Phys. Rev.*, **134**, A713 (1964).
17. D. Pugh, "Surface States on the $\langle 111 \rangle$ Surface of Diamond," *Phys. Rev. Lett.*, **12**, 390 (1964).

18. W. E. Spicer, P. W. Chye, C. M. Garner, I. Lindau, and P. Pianetta, "The Surface Electronic Structure of III-V Compounds and the Mechanism of Fermi Level Pinning by Oxygen (Passivation) and Metals (Schottky Barriers)," *Surface Sci.*, **86**, 763 (1979).
19. W. E. Spicer, I. Lindau, P. Skeath, C. Y. Su, and P. Chye, "Unified Mechanism for Schottky-Barrier Formation and III-V Oxide Interface States," *Phys. Rev. Lett.*, **44**, 420 (1980).
20. L. Pauling, *The Nature of The Chemical Bond*, 3rd Ed., Cornell University Press, Ithaca, New York, 1960.
21. S. Kurtin, T. C. McGill, and C. A. Mead, "Fundamental Transition in Electronic Nature of Solids," *Phys. Rev. Lett.*, **22**, 1433 (1969).
22. S. P. Murarka, *Silicides for VLSI Applications*, Academic Press, New York, 1983.
23. G. Ottaviani, K. N. Tu, and J. W. Mayer, "Interfacial Reaction and Schottky Barrier in Metal-Silicon Systems," *Phys. Rev. Lett.*, **44**, 284 (1980),
24. J. M. Andrews, *Extended Abstracts*, Electrochem. Soc. Spring Meet., Abstr. 191 (1975), p. 452.
25. S. M. Sze, C. R. Crowell, and D. Kahng, "Photoelectric Determination of the Image Force Dielectric Constant for Hot Electrons in Schottky Barriers," *J. Appl. Phys.*, **35**, 2534 (1964).
26. C. D. Salzberg and G. G. Villa, "Infrared Refractive Indexes of Silicon Germanium and Modified Selenium Glass," *J. Opt. Soc. Am.*, **47**, 244 (1957).
27. J. M. Shannon, "Reducing the Effective Height of a Schottky Barrier Using Low-Energy Ion Implantation," *Appl. Phys. Lett.*, **24**, 369 (1974).
28. J. M. Shannon, "Increasing the Effective Height of a Schottky Barrier Using Low-Energy Ion Implantation," *Appl. Phys. Lett.*, **25**, 75 (1974).
29. J. M. Andrews, R. M. Ryder, and S. M. Sze, "Schottky Barrier Diode Contacts," U.S. Patent 3,964,084 (1976).
30. J. M. Shannon, "Control of Schottky Barrier Height Using Highly Doped Surface Layers," *Solid-State Electron.*, **19**, 537 (1976).
31. C. R. Crowell, "The Richardson Constant for Thermionic Emission in Schottky Barrier Diodes," *Solid-State Electron.*, **8**, 395 (1965).
32. C. R. Crowell and S. M. Sze, "Current Transport in Metal-Semiconductor Barriers," *Solid-State Electron.*, **9**, 1035 (1966).
33. C. R. Crowell and S. M. Sze, "Electron-Optical-Phonon Scattering in the Emitter and Collector Barriers of Semiconductor-Metal-Semiconductor Structures," *Solid-State Electron.*, **8**, 979 (1965).
34. C. W. Kao, L. Anderson, and C. R. Crowell, "Photoelectron Injection at Metal-Semiconductor Interface," *Surface Sci.*, **95**, 321 (1980).
35. C. R. Crowell and S. M. Sze, "Quantum-Mechanical Reflection of Electrons at Metal-Semiconductor Barriers: Electron Transport in Semiconductor-Metal-Semiconductor Structures," *J. Appl. Phys.*, **37**, 2685 (1966).
36. C. Y. Chang and S. M. Sze, "Carrier Transport across Metal-Semiconductor Barriers," *Solid-State Electron.*, **13**, 727 (1970).
37. J. M. Andrews and M. P. Lepselter, "Reverse Current-Voltage Characteristics of Metal-Silicide Schottky Diodes," *Solid-State Electron.*, **13**, 1011 (1970).

38. C. R. Crowell and M. Beguwala, "Recombination Velocity Effects on Current Diffusion and Imref in Schottky Barriers," *Solid-State Electron.*, **14**, 1149 (1971).
39. F. A. Padovani and R. Stratton, "Field and Thermionic-Field Emission in Schottky Barriers," *Solid-State Electron.*, **9**, 695 (1966).
40. A. Y. C. Yu and E. H. Snow, "Minority Carrier Injection of Metal-Silicon Contacts," *Solid-State Electron.*, **12**, 155 (1969).
41. D. L. Scharfetter, "Minority Carrier Injection and Charge Storage in Epitaxial Schottky Barrier Diodes," *Solid-State Electron.*, **8**, 299 (1965).
42. H. C. Card, "Tunnelling MIS Structures," *Inst. Phys. Conf. Ser.*, **50**, 140 (1980).
43. M. Y. Doghish and F. D. Ho, "A Comprehensive Analytical Model for Metal-Insulator-Semiconductor (MIS) Devices," *IEEE Trans. Electron Dev.*, **ED-39**, 2771 (1992).
44. C. R. Crowell, J. C. Sarace, and S. M. Sze, "Tungsten-Semiconductor Schottky-Barrier Diodes," *Trans. Met. Soc. AIME*, **233**, 478 (1965).
45. M. P. Lepselter and S. M. Sze, "Silicon Schottky Barrier Diode with Near-Ideal *I-V* Characteristics," *Bell Syst. Tech. J.*, **47**, 195 (1968).
46. J. M. Andrews and F. B. Koch, "Formation of NiSi and Current Transport across the NiSi-Si Interface," *Solid-State Electron.*, **14**, 901 (1971).
47. K. Chino, "Behavior of Al-Si Schottky Barrier Diodes under Heat Treatment," *Solid-State Electron.*, **16**, 119 (1973).
48. A. M. Goodman, "Metal-Semiconductor Barrier Height Measurement by the Differential Capacitance Method—One Carrier System," *J. Appl. Phys.*, **34**, 329 (1963).
49. M. Beguwala and C. R. Crowell, "Characterization of Multiple Deep Level Systems in Semiconductor Junctions by Admittance Measurements," *Solid-State Electron.*, **17**, 203 (1974).
50. C. R. Crowell, W. G. Spitzer, L. E. Howarth, and E. Labate, "Attenuation Length Measurements of Hot Electrons in Metal Films," *Phys. Rev.*, **127**, 2006 (1962).
51. R. H. Fowler, "The Analysis of Photoelectric Sensitivity Curves for Clean Metals at Various Temperatures," *Phys. Rev.*, **38**, 45 (1931).
52. C. R. Crowell, S. M. Sze, and W. G. Spitzer, "Equality of the Temperature Dependence of the Gold-Silicon Surface Barrier and the Silicon Energy Gap in Au *n*-type Si Diodes," *Appl. Phys. Lett.*, **4**, 91 (1964).
53. J. 0. McCaldin, T. C. McGill, and C. A. Mead, "Schottky Barriers on Compound Semiconductors: The Role of the Anion," *J. Vac. Sci. Technol.*, **13**, 802 (1976).
54. J. M. Andrews, "The Role of the Metal-Semiconductor Interface in Silicon Integrated Circuit Technology," *J. Vac. Sci. Technol.*, **11**, 972 (1974).
55. A. G. Milnes, *Semiconductor Devices and Integrated Electronics*, Van Nostrand, New York, 1980.
56. *Properties of Silicon*, INSPEC, London, 1988.
57. *Properties of Gallium Arsenide*, INSPEC, London, 1986. 2nd Ed., 1996.
58. G. Myburg, F. D. Auret, W. E. Meyer, C. W. Louw, and M. J. van Staden, "Summary of Schottky Barrier Height Data on Epitaxially Grown *n*- and *p*-GaAs," *Thin Solid Films*, **325**, 181 (1998).
59. N. Newman, T. Kendelewicz, L. Bowman, and W. E. Spicer, "Electrical Study of Schottky Barrier Heights on Atomically Clean and Air-Exposed *n*-InP (110) Surfaces," *Appl. Phys. Lett.*, **46**, 1176 (1985).

60. J. M. Andrews and J. C. Phillips, "Chemical Bonding and Structure of Metal-Semiconductor Interfaces," *Phys. Rev. Lett.*, **35**, 56 (1975).
61. G. J. van Gurp, "The Growth of Metal Silicide Layers on Silicon," in H. R. Huff and E. Sirtl, Eds., *Semiconductor Silicon 1977*, Electrochemical Society, Princeton, New Jersey, 1977, p. 342.
62. I. Ohdomari, K. N. Tu, F. M. d'Heurle, T. S. Kuan, and S. Petersson, "Schottky-Barrier Height of Iridium Silicide," *Appl. Phys. Lett.*, **33**, 1028 (1978).
63. J. L. Saltich and L. E. Terry, "Effects of Pre- and Post-Annealing Treatments on Silicon Schottky Barrier Diodes," *Proc. IEEE*, **58**, 492 (1970).
64. A. K. Sinha, "Electrical Characteristics and Thermal Stability of Platinum Silicide-to-Silicon Ohmic Contacts Metallized with Tungsten," *J. Electrochem. Soc.*, **120**, 1767 (1973).
65. A. K. Sinha, T. E. Smith, M. H. Read, and J. M. Poate, "*n*-GaAs Schottky Diodes Metallized with Ti and Pt/Ti," *Solid-State Electron.*, **19**, 489 (1976).
66. Y. Itoh and N. Hashimoto, "Reaction-Process Dependence of Barrier Height between Tungsten Silicide and *n*-Type Silicon," *J. Appl. Phys.*, **40**, 425 (1969).
67. R. Tung, "Epitaxial Silicide Contacts," in R. Hull, Ed., *Properties of Crystalline Silicon*, INSPEC, London, 1999.
68. For general references on vacuum deposition, see L. Holland, *Vacuum Deposition of Thin Films*, Chapman & Hall, London, 1966; A. Roth, *Vacuum Technology*, North-Holland, Amsterdam, 1976.
69. R. E. Honig, "Vapor Pressure Data for the Solid and Liquid Elements," *RCA Rev.*, **23**, 567 (1962).
70. D. T. Young and J. C. Irvin, "Millimeter Frequency Conversion Using Au-*n*-type GaAs Schottky Barrier Epitaxy Diode with a Novel Contacting Technique," *Proc. IEEE.*, **53**, 2130 (1965).
71. D. Kahng and R. M. Ryder, "Small Area Semiconductor Devices," U.S. Patent 3,360,851 (1968).
72. A. Y. C. Yu and C. A. Mead, "Characteristics of Al-Si Schottky Barrier Diode," *Solid-State Electron.*, **13**, 97 (1970).
73. N. G. Anantha and K. G. Ashar, *IBM J. Res. Dev.*, **15**, 442 (1971).
74. C. Rhee, J. L. Saltich, and R. Zwernemann, "Moat-Etched Schottky Barrier Diode Displaying Near Ideal *I-V* Characteristics," *Solid-State Electron.*, **15**, 1181 (1972).
75. J. L. Saltich and L. E. Clark, "Use of a Double Diffused Guard Ring to Obtain Near Ideal *I-V* Characteristics in Schottky-Barrier Diodes," *Solid-State Electron.*, **13**, 857 (1970).
76. K. J. Linden, "GaAs Schottky Mixer Diode with Integral Guard Layer Structure," *IEEE Trans. Electron Dev.*, **ED-23**, 363 (1976).
77. A. Rusu, C. Bulucea, and C. Postolache, "The Metal-Overlap-Laterally-Diffused (MOLD) Schottky Diode," *Solid-State Electron.*, **20**, 499 (1977).
78. D. J. Coleman Jr., J. C, Irvin, and S. M. Sze, "GaAs Schottky Diodes with Near-Ideal Characteristics," *Proc. IEEE*, **59**, 1121 (1971).
79. K. Tada and J. L. R. Laraya, "Reduction of the Storage Time of a Transistor Using a Schottky-Barrier Diode," *Proc. IEEE*, **55**, 2064 (1967).

80. J. C. Irvin and N. C. Vanderwal, "Schottky-Barrier Devices," in H. A. Watson, Ed., *Microwave Semiconductor Devices and Their Circuit Applications*, McGraw-Hill, New York, 1968.
81. N. C. Vanderwal, "A Microwave Schottky-Barrier Varistor Using GaAs for Low Series Resistance," *Tech. Dig. IEEE IEDM*, (1967).
82. M. McColl and M. F. Millea, "Advantages of Mott Barrier Mixer Diodes," *Proc. IEEE*, **61**, 499 (1973).
83. C. Y. Chang, Y. K. Fang, and S. M. Sze, "Specific Contact Resistance of Metal-Semiconductor Barriers," *Solid-State Electron.*, **14**, 541 (1971).
84. A. Y. C. Yu, "Electron Tunneling arid Contact Resistance of Metal-Silicon Contact Barriers," *Solid-State Electron.*, **13**, 239 (1970).
85. C. R. Crowell and V. L. Rideout, "Normalized Thermionic-Field (T-F) Emission in Metal-Semiconductor (Schottky) Barriers," *Solid-State Electron.*, **12**, 89 (1969).
86. K. K. Ng and R. Liu, "On the Calculation of Specific Contact Resistivity on ⟨100⟩ Si," *IEEE Trans. Electron Dev.*, **ED-37**, 1535 (1990).
87. V. L. Rideout, "A Review of the Theory and Technology for Ohmic Contacts to Group III-V Compound Semiconductors," *Solid-State Electron.*, **18**, 541 (1975).
88. S. S. Li, *Semiconductor Physical Electronics*, Plenum Press, New York, 1993.
89. R. H. Cox and H. Strack, "Ohmic Contacts for GaAs Devices," *Solid-State Electron.*, **10**, 1213 (1967).
90. H. Murrmann and D. Widmann, "Current Crowding on Metal Contacts to Planar Devices," *IEEE Trans. Electron Dev.*, **ED-16**, 1022 (1969).

PROBLEMS

1. Draw the band diagrams of the conduction band and the Fermi level for GaAs metal-semiconductor contacts with n-type doping levels of (a) 10^{15} cm^{-3}, (b) 10^{17} cm^{-3}, and (c) 10^{18} cm^{-3}. The barrier height ($q\phi_{Bn0}$) is 0.80 eV.
2. For a Au-n-Si metal-semiconductor contact with a donor concentration of 2.8×10^{16} cm^{-3}, what is the Schottky-barrier lowering at thermal equilibrium and the corresponding location of the lowering. The barrier height ($q\phi_{Bn0}$) is 0.80 eV.
3. Derive Eq. 72. Fill in detailed steps in the derivation.
4. Find the minority current density and the injection ratio at a low-injection condition for a Au-Si Schottky-barrier diode with $\phi_{Bn} = 0.80$ V. The silicon is 1 Ω-cm, n-type with $\tau_p = 100$ μs.
5. Based on the theoretical result on p. 732 in the paper by Chang/Sze [*Solid-State Electron.*, **13**, 727 (1970)], find the ideality factor for a Schottky contact with $N_D = 10^{18}$ cm^{-3} at 77 K.
6. Derive Eq. 42 and find the limiting value of ϕ_B' for $p_1 > n_2$ and $ap_1 \gg Wn_2$.
7. The reverse saturation currents of a Schottky diode and a p-n junction at 300 K are 5×10^{-8} A and 10^{-12} A, respectively. The diodes are connected in series and are driven by a constant current of 0.5 mA. Find the total voltage across the diodes.
8. (a) Find the barrier height and donor concentration of the W-GaAs Schottky barrier shown in Fig. 30.

(b) Compare the barrier height with that obtained from the saturation current density of 5×10^{-7} A/cm^2 shown in Fig. 25, assume $A^{**} = 4$ A/cm^2-K^2.

(c) If there is a difference in barrier height, is the difference consistent with the Schottky-barrier lowering?

9. For a metal-n-Si contact, the barrier height obtained by photoelectric measurement is 0.65 V while the voltage intercept obtained from C-V measurement is 0.5 V. Find the doping concentration of the uniformly doped silicon substrate.

10. The capacitance of a Au-n-GaAs Schottky-barrier diode is given by the relation $1/C^2 = 1.57\times10^{15} - 2.12\times10^{15}\,V$, where C is expressed in μF and V is in volts. Taking the diode area to be 0.1 cm^2, calculate the built-in potential, the barrier height, and the dopant concentration.

11. The forward-bias cutoff frequency for a Pd-GaAs contact made on an n-type epitaxial layer of 0.5 μm thick is 370 GHz. If the circular contact area is 1.96×10^{-7} cm^2, find the depletion width under the forward-bias condition.

12. An ohmic contact has an area of 10^{-6} cm^2 and is formed on an n-type silicon with $N_D = 3\times10^{20}$ cm^{-3}. The barrier height ϕ_{Bn} is 0.8 V and the electron effective mass is $m_n^* = 0.26\,m_0$. Find the voltage drop across the contact when a forward current of 1 A flows through it.
{*Hint*: The current across the contact can be expressed as $I = I_0 \exp[-C_2(\phi_{Bn} - V)/\sqrt{N_D}]$ where I_0 is a constant and $C_2 \equiv 4\sqrt{m_n^* \varepsilon_s}/\hbar$.}

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

4

Metal-Insulator-Semiconductor Capacitors

4.1 INTRODUCTION
4.2 IDEAL MIS CAPACITOR
4.3 SILICON MOS CAPACITOR

4.1 INTRODUCTION

The metal-insulator-semiconductor (MIS) capacitor is the most useful device in the study of semiconductor surfaces. Since most practical problems in the reliability and stability of all semiconductor devices are intimately related to their surface conditions, an understanding of the surface physics with the help of MIS capacitors is of great importance to device operations. In this chapter we are concerned primarily with the metal-oxide-silicon (MOS) system. This system has been extensively studied because it is directly related to most silicon planar devices and integrated circuits.

The MIS structure was first proposed as a voltage-controlled varistor (variable capacitor) in 1959 by Moll[1] and by Pfann and Garrett.[2] Its characteristics were then analyzed by Frankl[3] and Lindner.[4] The first successful MIS structure was made of SiO_2 grown thermally on silicon surface by Ligenza and Spitzer in 1960.[5] This seminal experimental success immediately led to the first report of MOSFET by Kahng and Atalla.[6] Further study on this SiO_2-Si system was reported by Terman,[7] and Lehovec and Slobodskoy.[8] A comprehensive and in-depth treatment of the MOS capacitor can be found in *MOS Physics and Technology* by Nicollian and Brews.[9] The Si-SiO_2 system remains the most ideal and most practical MIS structure to date.

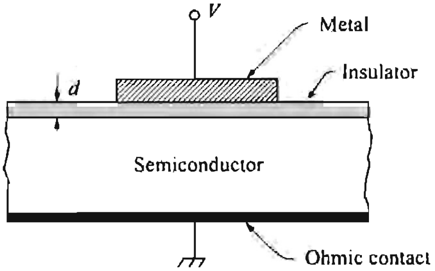

V
Metal
Insulator
d
Semiconductor
Ohmic contact

Fig. 1 Metal-insulator-semiconductor (MIS) capacitor, in its simplest form.

4.2 IDEAL MIS CAPACITOR

The metal-insulator-semiconductor (MIS) structure is shown in Fig. 1, where *d* is the thickness of the insulator and *V* is the applied voltage. Throughout this chapter we use the convention that the voltage *V* is positive when the metal plate is positively biased with respect to the semiconductor body.

The energy-band diagram of an ideal MIS structure without bias is shown in Fig. 2, for both *n*-type and *p*-type semiconductors. An ideal MIS capacitor is defined as follows: (1) The only charges that can exist in the structure under any biasing conditions are those in the semiconductor and those, with an equal but opposite sign, on the metal surface adjacent to the insulator, i.e., there is no interface trap nor any kind of oxide charge; (2) There is no carrier transport through the insulator under dc biasing conditions or the resistivity of the insulator is infinite. Furthermore, for the sake of simplicity we assume the metal is chosen such that the difference between the

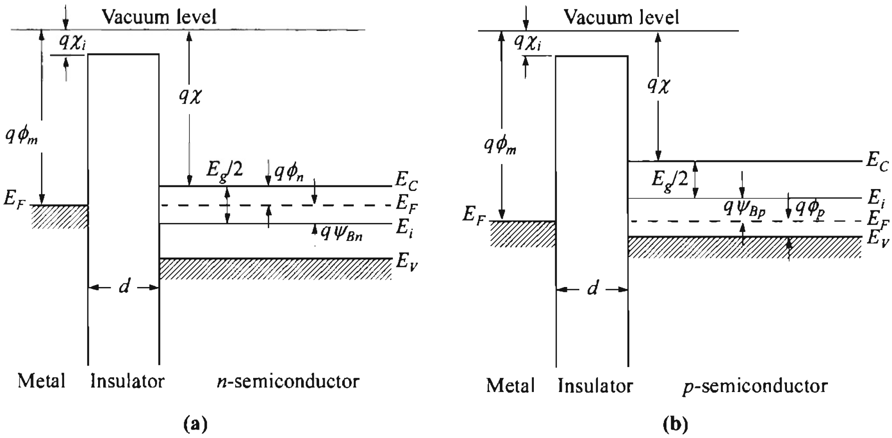

Vacuum level
$q\chi_i$
$q\chi$
$q\phi_m$
E_F
$E_g/2$
$q\phi_n$
E_C
E_F
E_i
$q\psi_{Bn}$
E_V
d
Metal
Insulator
n-semiconductor
(a)
Vacuum level
$q\chi_i$
$q\chi$
$q\phi_m$
E_F
$E_g/2$
$q\psi_{Bp}$
$q\phi_p$
E_C
E_i
E_F
E_V
d
Metal
Insulator
p-semiconductor
(b)

Fig. 2 Energy-band diagrams of ideal MIS capacitors at equilibrium ($V = 0$). (a) *n*-type semiconductor. (b) *p*-type semiconductor.

metal work function ϕ_m and the semiconductor work function is zero, or $\phi_{ms} = 0$. The above conditions, with the help of Fig. 2, are equivalent to:

$$\phi_{ms} \equiv \phi_m - \left(\chi + \frac{E_g}{2q} - \psi_{Bn}\right) = \phi_m - (\chi + \phi_n) = 0 \qquad \text{for } n\text{-type} \qquad (1a)$$

$$\phi_{ms} \equiv \phi_m - \left(\chi + \frac{E_g}{2q} + \psi_{Bp}\right) = \phi_m - \left(\chi + \frac{E_g}{q} - \phi_p\right) = 0 \qquad \text{for } p\text{-type} \qquad (1b)$$

where χ and χ_i are the electron affinities for the semiconductor and insulator respectively, and ψ_{Bn}, ψ_{Bp}, ϕ_n, ϕ_p are the Fermi potentials with respect to the midgap and band edges. In other words, the band is flat (flat-band condition) when there is no applied voltage. The ideal MIS capacitor theory to be considered in this section serves as a foundation for understanding practical MIS structures and to exploring the physics of semiconductor surfaces.

When an ideal MIS capacitor is biased with positive or negative voltages, basically three cases may exist at the semiconductor surface (Fig. 3). Consider the p-type semiconductor first (top figures). When a negative voltage ($V < 0$) is applied to the metal plate, the valence-band edge E_V bends upward near the surface and is closer to the Fermi level (Fig. 3a). For an ideal MIS capacitor, no current flows in the structure (or $dE_F/dx = 0$), so the Fermi level remains flat in the semiconductor. Since the carrier

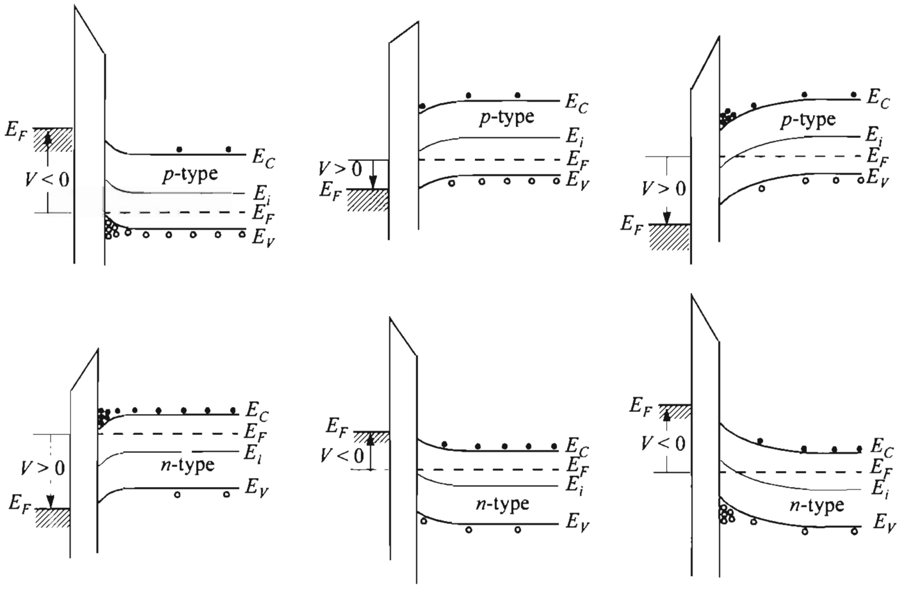


Fig. 3 Energy-band diagrams for ideal MIS capacitors under different bias, for the conditions of: (a) accumulation, (b) depletion, and (c) inversion. Top/bottom figures are for p-type/n-type semiconductor substrates.

density depends exponentially on the energy difference ($E_F - E_V$), this band bending causes an accumulation of majority carriers (holes) near the semiconductor surface. This is the *accumulation* case. When a small positive voltage ($V > 0$) is applied, the bands bend downward, and the majority carriers are depleted (Fig. 3b). This is the *depletion* case. When a larger positive voltage is applied, the bands bend even more downward so that the intrinsic level E_i at the surface crosses over the Fermi level E_F (Fig. 3c). At this point the number of electrons (minority carriers) at the surface is larger than that of the holes, the surface is thus inverted and this is the *inversion* case. Similar results can be obtained for the *n*-type semiconductor. The polarity of the voltage, however, should be changed for the *n*-type semiconductor.

4.2.1 Surface Space-Charge Region

In this section we derive the relations between the surface potential, space charge, and electric field. These relations are then used to derive the capacitance-voltage characteristics of the ideal MIS structure in the following section.

Figure 4 shows a more detailed band diagram at the surface of a *p*-type semiconductor. The potential $\psi_p(x)$ is defined as the potential $E_i(x)/q$ with respect to the bulk of the semiconductor;

$$\psi_p(x) \equiv -\frac{[E_i(x) - E_i(\infty)]}{q}. \tag{2}$$

At the semiconductor surface, $\psi_p(0) \equiv \psi_s$, and ψ_s is called the surface potential. The electron and hole concentrations as a function of ψ_p are given by the following relations:

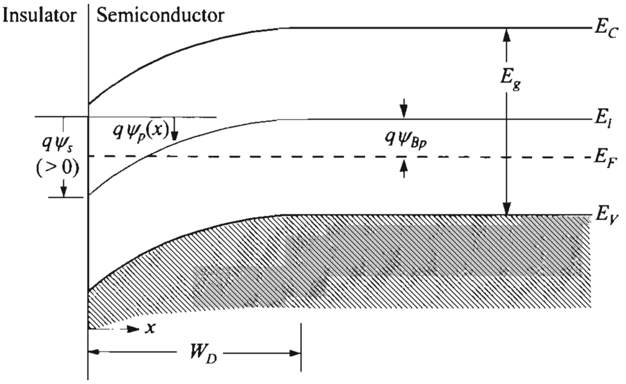


Fig. 4 Energy-band diagram at the surface of a *p*-type semiconductor. The potential energy $q\psi_p$ is measured with respect to the intrinsic Fermi level E_i in the bulk. The surface potential ψ_s is positive as shown. Accumulation occurs when $\psi_s < 0$. Depletion occurs when $\psi_{Bp} > \psi_s > 0$. Inversion occurs when $\psi_s > \psi_{Bp}$.

$$n_p(x) = n_{po}\exp\left(\frac{q\psi_p}{kT}\right) = n_{po}\exp(\beta\psi_p) \tag{3a}$$

$$p_p(x) = p_{po}\exp\left(\frac{-q\psi_p}{kT}\right) = p_{po}\exp(-\beta\psi_p) \tag{3b}$$

where ψ_p is positive when the band is bent downward (as shown in Fig. 4), n_{po} and p_{po} are the equilibrium densities of electrons and holes, respectively, in the bulk of the semiconductor, and $\beta \equiv q/kT$. At the surface the densities are

$$n_p(0) = n_{po}\exp(\beta\psi_s), \tag{4a}$$

$$p_p(0) = p_{po}\exp(-\beta\psi_s). \tag{4b}$$

From previous discussions and with the help of the above equations, the following regions of surface potential can be distinguished:

$\psi_s < 0$	Accumulation of holes (bands bending upward).
$\psi_s = 0$	Flat-band condition.
$\psi_{Bp} > \psi_s > 0$	Depletion of holes (bands bending downward).
$\psi_s = \psi_{Bp}$	Fermi-level at midgap, $E_F = E_i(0)$, $n_p(0) = p_p(0) = n_i$.
$2\psi_{Bp} > \psi_s > \psi_{Bp}$	Weak inversion [electron enhancement, $n_p(0) > p_p(0)$].
$\psi_s > 2\psi_{Bp}$	Strong inversion [$n_p(0) > p_{po}$ or N_A].

The potential $\psi_p(x)$ as a function of distance can be obtained by using the one-dimensional Poisson equation

$$\frac{d^2\psi_p}{dx^2} = -\frac{\rho(x)}{\varepsilon_s}, \tag{5}$$

where $\rho(x)$ is the total space-charge density given by

$$\rho(x) = q(N_D^+ - N_A^- + p_p - n_p), \tag{6}$$

N_D^+ and N_A^- are the densities of the ionized donors and acceptors, respectively. Now, in the bulk of the semiconductor, far from the surface, charge neutrality must exist. Therefore at $\psi_p(\infty) = 0$, we have $\rho(x) = 0$ and

$$N_D^+ - N_A^- = n_{po} - p_{po} \tag{7}$$

The resultant Poisson equation to be solved within the depletion region is therefore

$$\begin{aligned}\frac{d^2\psi_p}{dx^2} &= -\frac{q}{\varepsilon_s}(n_{po} - p_{po} + p_p - n_p)\\ &= -\frac{q}{\varepsilon_s}\{p_{po}[\exp(-\beta\psi_p) - 1] - n_{po}[\exp(\beta\psi_p) - 1]\}.\end{aligned} \tag{8}$$

Integrating Eq. 8 from the surface toward the bulk[10]

$$\int_0^{d\psi_p/dx}\left(\frac{d\psi_p}{dx}\right)d\left(\frac{d\psi_p}{dx}\right) = \frac{-q}{\varepsilon_s}\int_0^{\psi_p}\{p_{po}[\exp(-\beta\psi_p) - 1] - n_{po}[\exp(\beta\psi_p) - 1]\}d\psi_p \tag{9}$$

gives the relation between the electric field ($\mathscr{E} \equiv -d\psi_p/dx$) and the potential ψ_p:

$$\mathscr{E}^2 = \left(\frac{2kT}{q}\right)^2\left(\frac{qp_{po}\beta}{2\varepsilon_s}\right)\left\{[\exp(-\beta\psi_p)+\beta\psi_p-1]+\frac{n_{po}}{p_{po}}[\exp(\beta\psi_p)-\beta\psi_p-1]\right\}. \quad (10)$$

We shall use the following abbreviations:

$$L_D \equiv \sqrt{\frac{kT\varepsilon_s}{p_{po}q^2}} \equiv \sqrt{\frac{\varepsilon_s}{qp_{po}\beta}}, \quad (11)$$

$$F\left(\beta\psi_p, \frac{n_{po}}{p_{po}}\right) \equiv \sqrt{[\exp(-\beta\psi_p)+\beta\psi_p-1]+\frac{n_{po}}{p_{po}}[\exp(\beta\psi_p)-\beta\psi_p-1]} \geq 0, \quad (12)$$

where L_D is the extrinsic Debye length for holes. [Note that $n_{po}/p_{po} = \exp(-2\beta\psi_{Bp})$.] Thus the electric field is given by

$$\mathscr{E}(x) = \pm\frac{\sqrt{2}kT}{qL_D}F\left(\beta\psi_p, \frac{n_{po}}{p_{po}}\right), \quad (13)$$

with positive sign for $\psi_p > 0$ and negative sign for $\psi_p < 0$. To determine the electric field at the surface $\mathscr{E}_s$, we let $\psi_p = \psi_s$:

$$\mathscr{E}_s = \pm\frac{\sqrt{2}kT}{qL_D}F\left(\beta\psi_s, \frac{n_{po}}{p_{po}}\right). \quad (14)$$

From this surface field, we can deduce the total space charge per unit area by applying Gauss' law:

$$Q_s = -\varepsilon_s\mathscr{E}_s = \mp\frac{\sqrt{2}\varepsilon_s kT}{qL_D}F\left(\beta\psi_s, \frac{n_{po}}{p_{po}}\right). \quad (15)$$

A typical variation of the space-charge density Q_s as a function of the surface potential ψ_s is shown in Fig. 5, for a p-type silicon with $N_A = 4\times10^{15}$ cm^{-3} at room temperature. Note that for negative ψ_s, Q_s is positive and it corresponds to the accumulation region. The function F is dominated by the first term in Eq. 12, that is, $Q_s \propto \exp(q|\psi_s|/2kT)$. For $\psi_s = 0$, we have the flat-band condition and $Q_s = 0$. For $2\psi_B > \psi_s > 0$, Q_s is negative and we have the depletion and weak-inversion cases. The function F is now dominated by the second term, that is, $Q_s \propto \sqrt{\psi_s}$. For $\psi_s > 2\psi_B$, we have the strong inversion case with the function F dominated by the fourth term, that is, $Q_s \propto \exp(q\psi_s/2kT)$. Also note that this strong inversion begins at a surface potential,

$$\psi_s(\text{strong inversion}) \approx 2\psi_{Bp} \approx \frac{2kT}{q}\ln\left(\frac{N_A}{n_i}\right). \quad (16)$$

4.2.2 Ideal MIS Capacitance Curves

Figure 6a shows the band diagram of an ideal MIS structure with the band bending of the semiconductor similar to that shown in Fig. 4 but in strong inversion. The charge distribution is shown in Fig. 6b. For charge neutrality of the system, it is required that

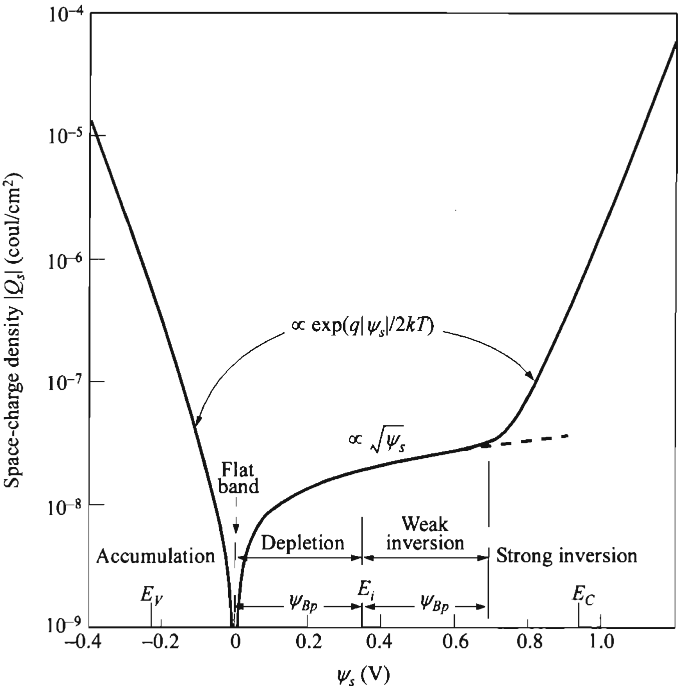


Fig. 5 Variation of space-charge density in the semiconductor as a function of the surface potential ψ_s, for a p-type silicon with $N_A = 4\times10^{15}$ cm^{-3} at room temperature.

$$Q_M = -(Q_n + qN_A W_D) = -Q_s \tag{17}$$

where Q_M is charges per unit area on the metal, Q_n is the electrons per unit area near the surface of the inversion region, $qN_A W_D$ is the ionized acceptors per unit area in the space-charge region with depletion width W_D, and Q_s is the total charges per unit area in the semiconductor. The electric field and the potential as obtained by first and second integrations of the Poisson equation are shown in Figs. 6c and d, respectively.

Clearly, in the absence of any work-function difference, the applied voltage will partly appear across the insulator and partly across the semiconductor. Thus

$$V = V_i + \psi_s \tag{18}$$

where V_i is the potential across the insulator and is given (Fig. 6c) by

$$V_i = \mathscr{E}_i d = \frac{|Q_s|d}{\varepsilon_i} = \frac{|Q_s|}{C_i}. \tag{19}$$

The total capacitance C of the system is a series combination of the insulator capacitance

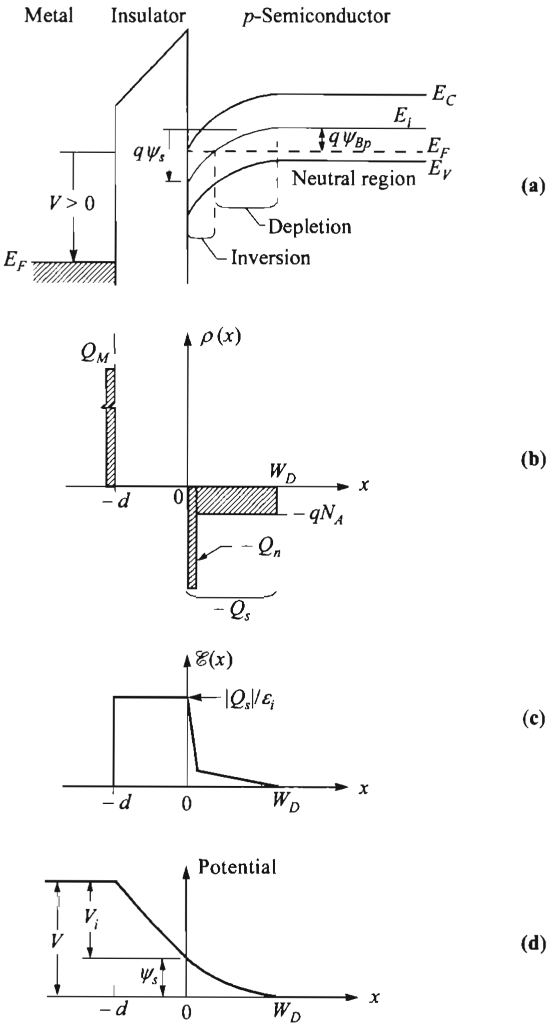

Metal
Insulator
p-Semiconductor
E_C
E_i
$q\psi_s$
$q\psi_{Bp}$
E_F
E_V
Neutral region
$V > 0$
E_F
Depletion
Inversion
(a)
Q_M
$\rho(x)$
W_D
x
$-d$
0
$-qN_A$
$-Q_n$
$-Q_s$
(b)
$\mathscr{E}(x)$
$|Q_s|/\varepsilon_i$
x
$-d$
0
W_D
(c)
Potential
V_i
V
ψ_s
x
$-d$
0
W_D
(d)

Fig. 6 (a) Band diagram of an ideal MIS capacitor under strong inversion. (b) Charge distribution. (c) Electric-field distribution. (d) Potential distribution (relative to the semiconductor bulk).

$$C_i = \frac{\varepsilon_i}{d}, \tag{20}$$

and the semiconductor depletion-layer capacitance C_D:

$$C = \frac{C_i C_D}{C_i + C_D}. \tag{21}$$

For a given insulator thickness d, the value of C_i is constant and corresponds to the maximum capacitance of the system. But the semiconductor capacitance C_D not only depends on the bias (or ψ_s), it is also a function of the measurement frequency. Figure 7 illustrates the vastly different characteristics of C-V curves measured at different frequencies and sweep rates. The difference mainly occurs at the inversion regime, especially strong inversion. Figure 7 also shows the corresponding surface potentials at different regimes. For an ideal MIS capacitor (and without work-function difference), flat-band occurs at $V = 0$, where $\psi_s = 0$. The depletion regime corresponds to a surface potential ranging from $\psi_s = 0$ to $\psi_s = \psi_{Bp}$. Weak inversion begins at $\psi_s = \psi_{Bp}$, and the onset of strong inversion occurs at $\psi_s = 2\psi_{Bp}$. The minimum low-frequency capacitance $C_{\min}$ occurs inbetween these two points.

Low-Frequency Capacitance. The capacitance of the semiconductor depletion layer is obtained by differentiating the total static charge in the semiconductor side (Eq. 15) with respect to the semiconductor surface potential,

$$C_D \equiv \frac{dQ_s}{d\psi_s} = \frac{\varepsilon_s}{\sqrt{2}L_D} \frac{1 - \exp(-\beta\psi_s) + (n_{po}/p_{po})[\exp(\beta\psi_s) - 1]}{F(\beta\psi_s, n_{po}/p_{po})}. \tag{22}$$

This capacitance can be visualized as the slope in Fig. 5. Combination of Eqs. 18–22 gives the complete description of the ideal low-frequency C-V curve as shown in Fig. 7 curve (a).

In describing this low-frequency curve we begin at the left side (negative voltage and ψ_s), where we have an accumulation of holes and therefore a high differential capacitance of the semiconductor. As a result the total capacitance is close to the insulator capacitance. As the negative voltage is reduced to zero, we have the flat-band condition, that is, $\psi_s = 0$. Since the function F approaches zero, C_D has to be obtained from Eq. 22 by expanding the exponential terms into series, and we obtain

$$C_D(\text{flat-band}) = \frac{\varepsilon_s}{L_D}. \tag{23}$$

The total capacitance at flat-band condition is given by Eqs. 21 and 23,

$$C_{FB}(\psi_s = 0) = \frac{\varepsilon_i \varepsilon_s}{\varepsilon_s d + \varepsilon_i L_D} = \frac{\varepsilon_i \varepsilon_s}{\varepsilon_s d + \varepsilon_i \sqrt{kT\varepsilon_s / N_A q^2}} \tag{24}$$

where ε_i and ε_s are the permittivities of the insulator and the semiconductor respectively, and L_D is the extrinsic Debye length given by Eq. 11.

It can be shown that under depletion and weak inversion conditions, i.e. $2\psi_{Bp} > \psi_s > kT/q$, the function F (Eq. 12) can be simplified to

$$F \approx \sqrt{\beta\psi_s} \qquad (2\psi_{Bp} > \psi_s > kT/q). \tag{25}$$

With this, the space-charge density (Eq. 15) can be reduced to

$$Q_s = \sqrt{2\varepsilon_s q p_{po} \psi_s} = qW_D N_A \qquad (2\psi_{Bp} > \psi_s > kT/q) \tag{26}$$

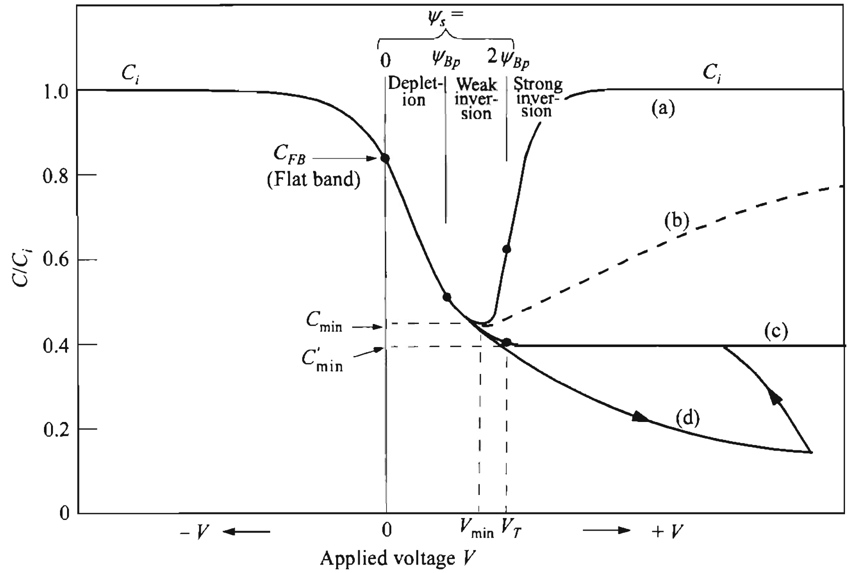


Fig. 7 MIS C-V curves. Voltage is applied to the metal relative to the p-semiconductor. (a) Low frequency. (b) Intermediate frequency. (c) High frequency. (d) High frequency with fast sweep (deep depletion). Flat-band voltage of $V = 0$ is assumed.

which is the familiar depletion approximation. From Eqs. 18, 19, and 26, we can express the depletion width as a function of the terminal voltages. The quadratic equation gives a solution of

$$W_D = \sqrt{\frac{\varepsilon_s^2}{C_{ox}^2} + \frac{2\varepsilon_s V}{qN_D}} - \frac{\varepsilon_s}{C_{ox}}. \tag{27}$$

Once W_D is known, C_D and ψ_s are deduced. The depletion capacitance (Eq. 22) can be estimated by

$$C_D = \sqrt{\frac{\varepsilon_s q p_{po}}{2\psi_s}} = \frac{\varepsilon_s}{W_D} \qquad (2\psi_{Bp} > \psi_s > kT/q). \tag{28}$$

With further increase in positive voltage, the depletion region widens which acts as a dielectric at the semiconductor surface in series with the insulator, and the total capacitance continues to decrease. The capacitance goes through a minimum and then increases again as the inversion layer of electrons forms at the surface. The minimum capacitance and the corresponding minimum voltage are designated C_{min} and V_{min} respectively (Fig. 7). Since C_i is fixed, C_{min} can be found by the minimum value of C_D. The value of ψ_s corresponding to the minimum C_D can be obtained by differentiation of Eq. 22 and setting it to zero, resulting in a transcendental equation[9]

$$\sqrt{\cosh(\beta\psi_s - \beta\psi_B)} = \frac{\sinh(\beta\psi_s - \beta\psi_B) - \sinh(-\beta\psi_B)}{\sqrt{N_A/n_i}F(\beta\psi_s, n_{po}/p_{po})}. \tag{29}$$

With a known ψ_s, $C_{\min}$ and $V_{\min}$ can be determined from Eqs. 18–22.

Note that the increase of the capacitance depends on the ability of the electron concentration to follow the applied ac signal. This only happens at low frequencies where the recombination-generation rates of minority carriers (in our example, electrons) can keep up with the small-signal variation and lead to charge exchange with the inversion layer in step with the measurement signal. Unlike depletion and weak inversion, at strong inversion the incremental charge is no longer at the edge of the depletion region but at the semiconductor surface inversion layer, resulting in a large capacitance. The placement of the incremental charge at the semiconductor side is depicted in Fig. 8 for the different cases of low frequency, high frequency, and deep depletion. Experimentally, it is found that for the metal-SiO$_2$-Si system the range in which the capacitance is most frequency-dependent is between 5 Hz and 1 kHz.[11,12] This is related to the carrier lifetime and thermal generation rate in the silicon substrate. As a consequence, MOS curves measured at higher frequencies do not show the increase of capacitance in strong inversion, Fig. 7 curve (c).

High-Frequency Capacitance. The high-frequency curve can be obtained using an approach analogous to a one-sided abrupt *p-n* junction.[13,14] When the semiconductor surface is depleted, the ionized acceptors in the depletion region are given by $-qN_AW_D$, where W_D is the depletion width. Integrating the Poisson equation yields the potential distribution in the depletion region:

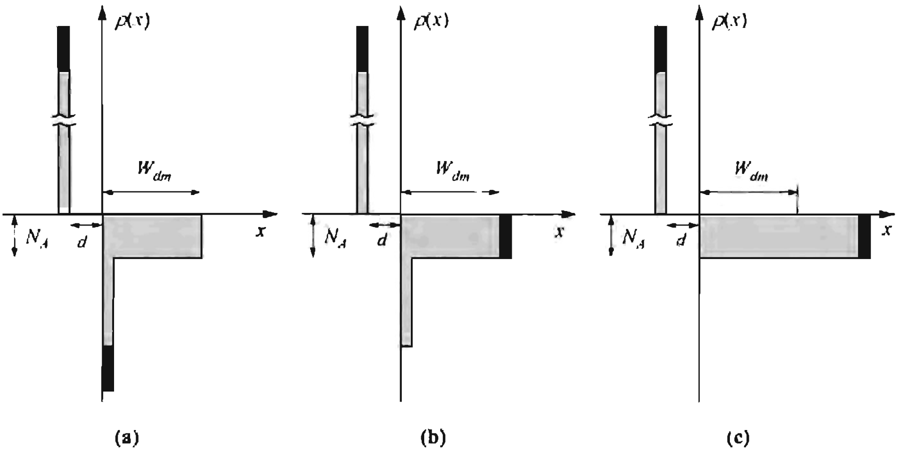


Fig. 8 In strong inversion, capacitance is a function of the small-signal frequency and the quiescent sweep rate. The incremental displacement charge (black area) is shown in cases of (a) low frequency, (b) high frequency, and (c) high frequency with a fast sweep rate (deep depletion, $W_D > W_{Dm}$).

$$\psi_p(x) = \psi_s\left(1 - \frac{x}{W_D}\right)^2 \tag{30}$$

where the surface potential ψ_s is given by

$$\psi_s = \frac{qN_A W_D^2}{2\varepsilon_s}. \tag{31}$$

When the applied voltage increases, ψ_s and W_D increase. Eventually, strong inversion will occur. As shown in Fig. 5, strong inversion begins at $\psi_s \approx 2\psi_B$. Once strong inversion occurs, the depletion-layer width reaches a maximum. When the bands are bent down far enough that $\psi_s = 2\psi_B$, the semiconductor is effectively shielded from further penetration of the electric field by the inversion layer and even a very small increase in band bending (corresponding to a very small increase in the depletion-layer width) results in a very large increase in the charge density within the inversion layer. Accordingly, the maximum width W_{Dm} of the depletion region under steady-state condition can be obtained from Eq. 16,

$$W_{Dm} \approx \sqrt{\frac{2\varepsilon_s \psi_s(\text{strong inv})}{qN_A}} \approx \sqrt{\frac{4\varepsilon_s kT\ln(N_A/n_i)}{q^2 N_A}}. \tag{32}$$

The relationship between W_{Dm} and the impurity concentration is shown in Fig. 9 for Si and GaAs, where N is equal to N_A for p-type and N_D for n-type semiconductors. This phenomena of maximum depletion width is unique to the MIS structure, and it does not occur in p-n junctions or Schottky barriers.

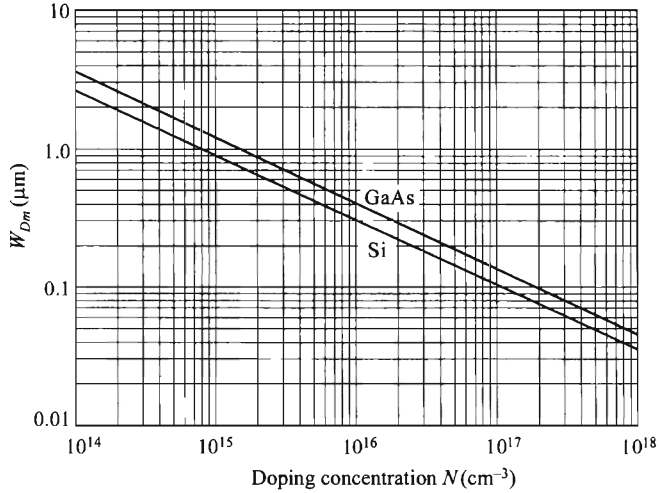


Fig. 9 Maximum depletion-layer width versus impurity concentration of semiconductors Si and GaAs under a heavy-inversion condition.

Another quantity of interest is the so-called turn-on voltage or threshold voltage, V_T, at which strong inversion occurs. From Eq. 18 and with proper substitutions, we obtain

$$V_T = \frac{|Q_s|}{C_i} + 2\psi_{Bp}$$
$$= \frac{\sqrt{2\varepsilon_s q N_A (2\psi_{Bp})}}{C_i} + 2\psi_{Bp}. \tag{33}$$

Note that even though the slow-varying quiescent voltage puts the additional charge at the surface inversion layer, the high-frequency small signal is too fast for the minority carriers and the incremental charge is put at the edge of the depletion region, as shown in Fig. 8b. The depletion capacitance is simply given by ε_i/W_D, with a minimum value corresponding to the maximum depletion width W_{DM}

$$C'_{\min} = \frac{\varepsilon_i \varepsilon_s}{\varepsilon_s d + \varepsilon_i W_{Dm}}. \tag{34}$$

Complete ideal C-V curves of the metal-SiO_2-Si system have been computed for various oxide thicknesses and semiconductor doping densities.[15] Figure 10a shows typical ideal C-V curves for p-type silicon. Note that as the oxide film becomes thinner, larger variation of the capacitance is obtained. Also the curves are sharper, reducing the threshold voltage V_T. Figure 10b shows the dependence of ψ_s on the applied voltage for the same systems. Similarly, modulation of ψ_s is more effective with thinner oxides.

The critical parameters C_{FB}, $C_{\min}$, $C'_{\min}$, V_T, and $V_{\min}$ are calculated and plotted in Fig. 11. These ideal MIS curves will be used in subsequent sections to compare with experimental results and to understand practical MIS systems. The conversion to n-type silicon is achieved simply by changing the sign of the voltage axes. Converting to other insulators requires scaling the oxide thickness with the ratio of the permittivities of SiO_2 and the other insulator

$$d_c = d_i \frac{\varepsilon_i(\mathrm{SiO_2})}{\varepsilon_i(\mathrm{insulator})} \tag{35}$$

where d_c is the equivalent SiO_2 thickness to be used in these curves, d_i and ε_i are the thickness and permittivity of the new insulator. For other semiconductors, the MIS curves similar to those in Fig. 10 can be constructed by using Eqs. 24 through 33.

At high frequency and with a fast sweeping ramp in the direction toward strong inversion, the semiconductor does not have enough time to come to equilibrium even with the large-signal variation. Deep depletion is said to occur when the depletion width is wider than the maximum value at equilibrium. This is the condition which CCDs are operated under when they are driven with large bias pulses, to be discussed in Section 13.6. The depletion width and the incremental charge are shown in Fig. 8c for comparison. Figure 7 curve (d) shows that the capacitance will keep on decreasing with bias, which is similar to a p-n junction or Schottky barrier. At even higher voltages, impact ionization can occur in the semiconductor, to be discussed later in con-

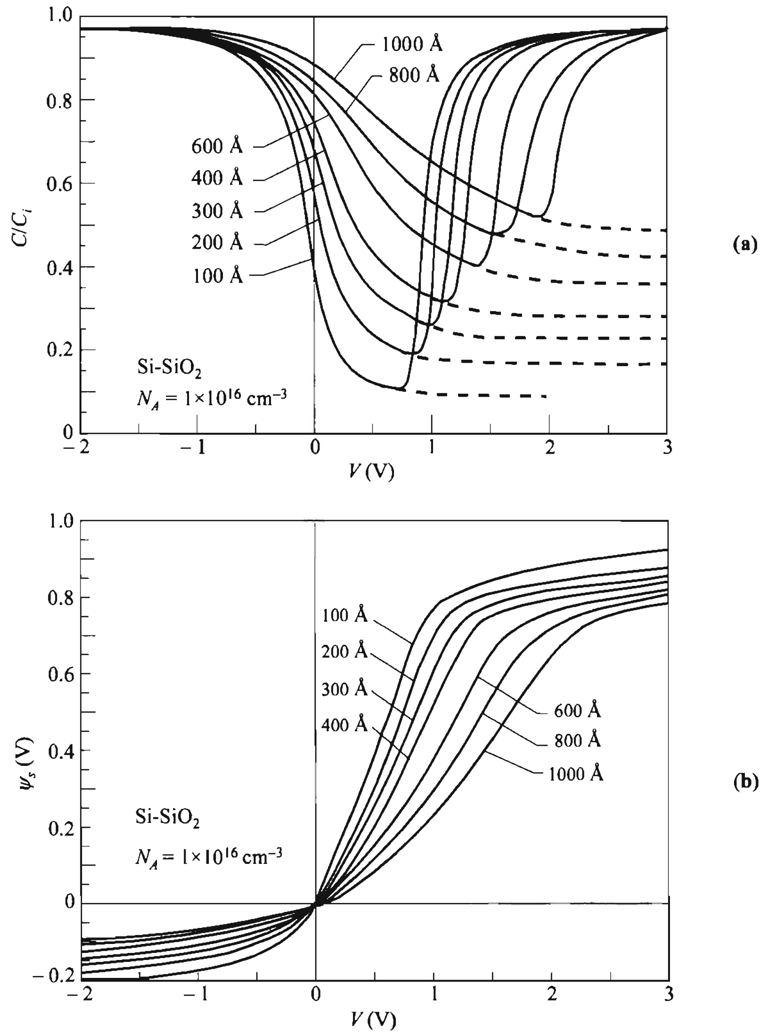


Fig. 10 (a) Ideal MOS C-V curves for various oxide thickness. Solid lines for low frequencies. Dashed lines for high frequencies. (b) Surface potential ψ_s vs. applied voltage. (After Ref. 15.)

nection with the avalanche effect. Under light illumination (see Section 4.3.5), however, extra minority carriers can be generated quickly and curve (d) will collapse to curve (c).

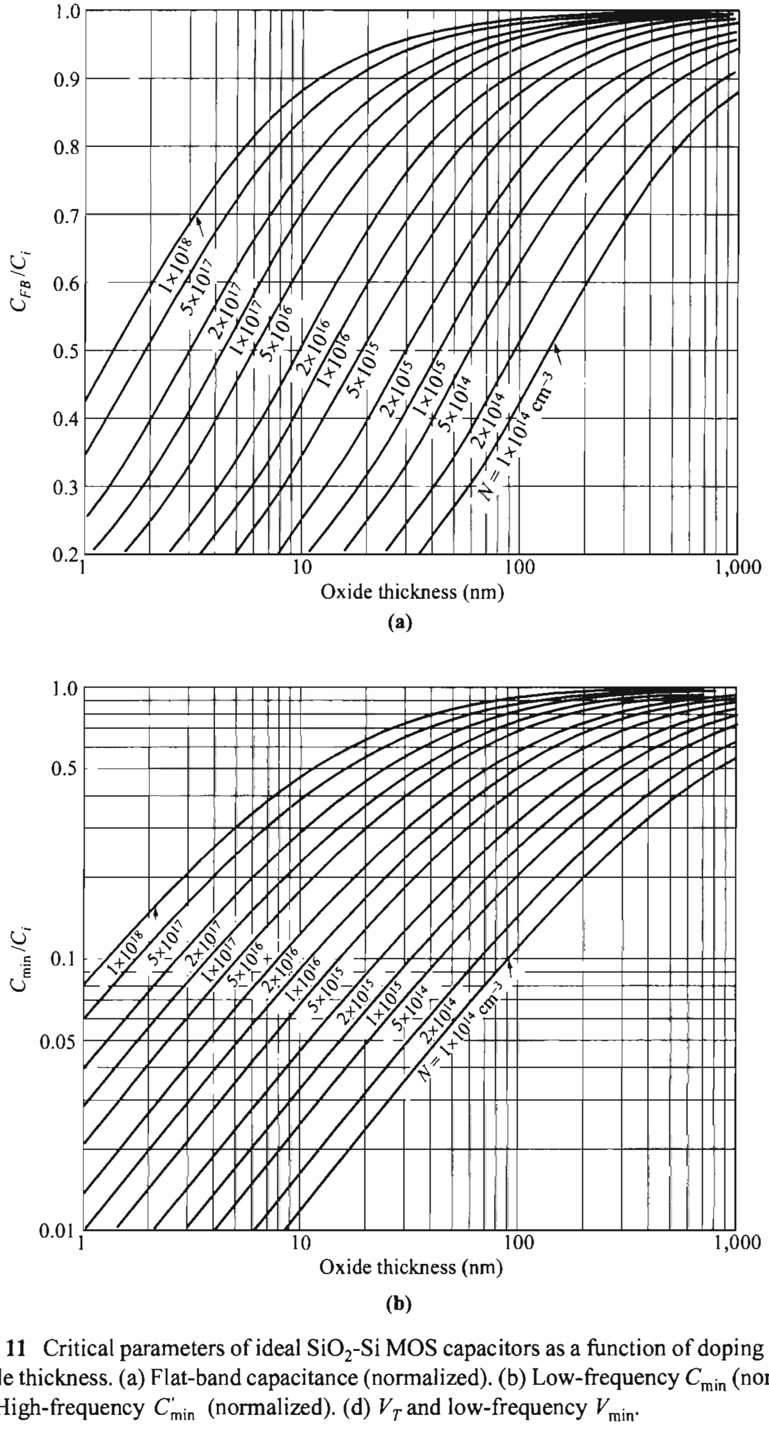
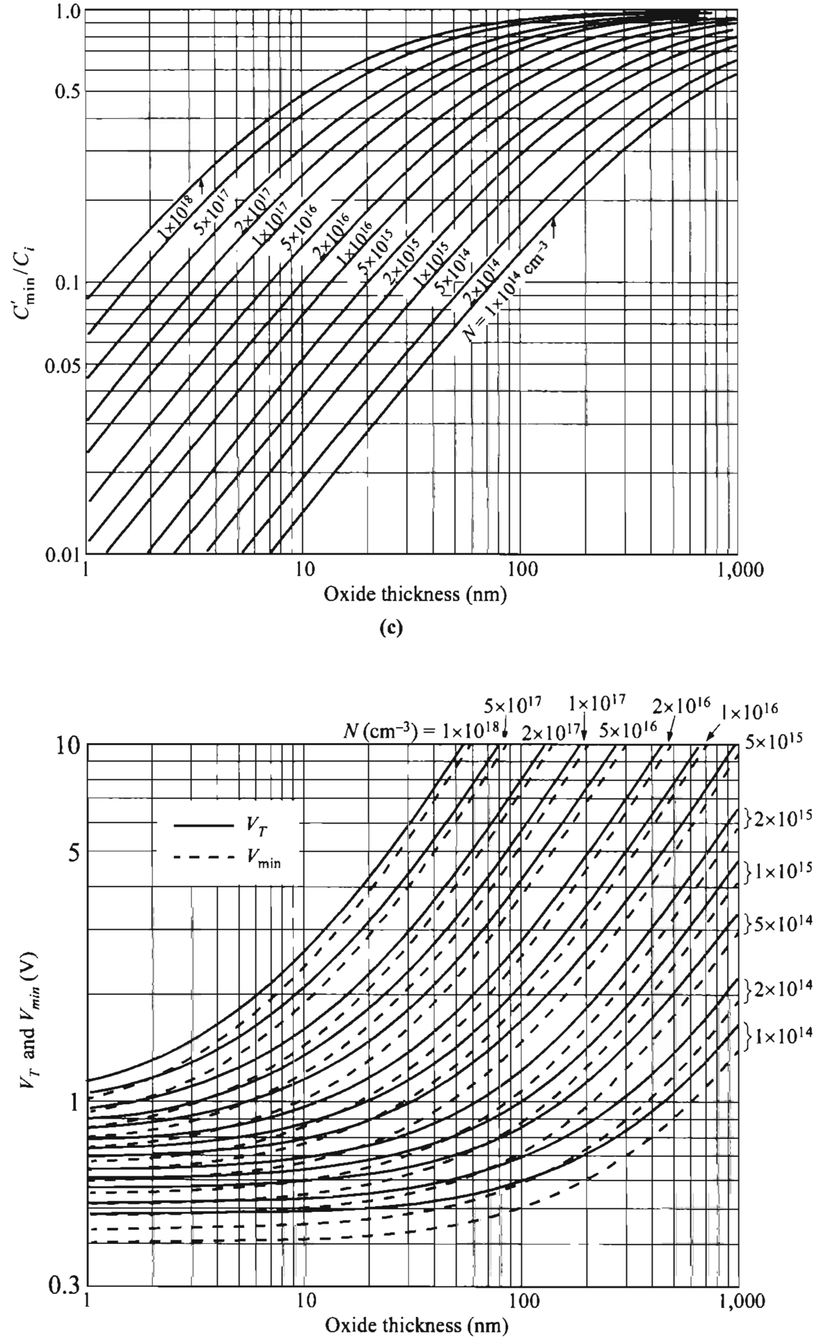


Fig. 11 Critical parameters of ideal SiO_2-Si MOS capacitors as a function of doping level and oxide thickness. (a) Flat-band capacitance (normalized). (b) Low-frequency C_{min} (normalized) (c) High-frequency C'_{min} (normalized). (d) V_T and low-frequency V_{min}.



(c)

(d)

Fig. 11 (Continued)

4.3 SILICON MOS CAPACITOR

Of all the MIS capacitors, the metal-oxide-silicon (MOS) capacitor is by far the most practical and important, and is used as an example here. An appealing picture of the interface is the chemical composition of the interfacial regions as a consequence of thermal oxidation.[9] It consists of a single-crystal silicon followed by a monolayer of SiO_x, that is, incompletely oxidized silicon, then a thin strained region of SiO_2, and the remainder stoichiometric, strain-free, amorphous SiO_2. (The compound SiO_x is stoichiometric when $x = 2$ and nonstoichiometric when $2 > x > 1$). For a practical MOS capacitor, interface traps and oxide charges exist that will, in one way or another, affect the ideal MOS characteristics.

The basic classifications of these traps and charges are shown in Fig. 12: (1) Interface traps of density D_{it} and trapped charges Q_{it}, which are located at the Si-SiO_2 interface with energy states within the silicon forbidden bandgap and which can exchange charges with silicon in a short time; Q_{it} is also determined by the occupancy or the Fermi level so its amount is bias dependent. Interface traps can possibly be produced by excess silicon (trivalent silicon), broken Si-H bonds, excess oxygen and impurities. (2) Fixed oxide charges Q_f, which are located at or near the interface and are immobile under an applied electric field. (3) Oxide trapped charges Q_{ot}, which can be created, for example, by X-ray radiation or hot-electron injection; these traps are distributed inside the oxide layer. (4) Mobile ionic charges Q_m, such as sodium ions, which are mobile within the oxide under bias-temperature stress conditions.

4.3.1 Interface Traps

Tamm,[17] Shockley,[18] and others[9] have studied the charge Q_{it} in interface traps (historically also called interface states, fast states, or surface states) and have shown that Q_{it} exists within the forbidden gap due to the interruption of the periodic lattice structure

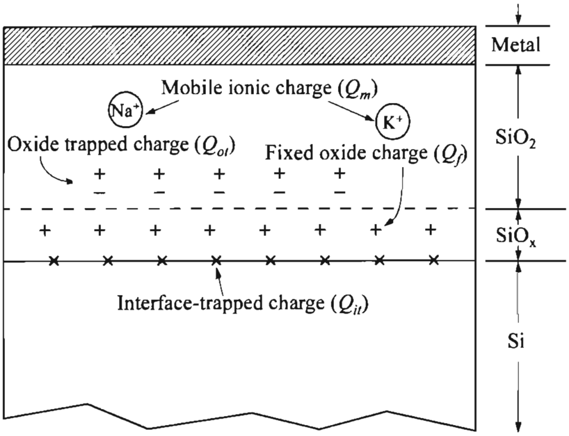

Metal
Mobile ionic charge (Q_m)
Na^+
K^+
SiO_2
Oxide trapped charge (Q_{ot})
Fixed oxide charge (Q_f)
SiO_x
Interface-trapped charge (Q_{it})
Si

Fig. 12 Terminology for charges associated with thermally oxidized silicon. (After Ref. 16.)

at the surface of a crystal. Shockley and Pearson experimentally found the existence of Q_{it} in their surface conductance measurement.[19] Measurements on clean surfaces[20] in an ultra-high-vacuum system confirm that Q_{it} can be very high—on the order of the density of surface atoms ($\approx 10^{15}$ atoms/cm^2). For the present MOS capacitors having thermally grown SiO_2 on Si, most of the interface-trapped charge can be neutralized by low-temperature (450°C) hydrogen annealing. The total surface traps can be as low as 10^{10} cm^{-2}, which amounts to about one interface trap per 10^5 surface atoms.

Similar to bulk impurities, an interface trap is considered a donor if it is neutral and can become positively charged by donating (giving up) an electron. An acceptor interface trap is neutral and becomes negatively charged by accepting an electron. The distribution functions (occupancy) for the interface traps are similar to those for the bulk impurity levels as discussed in Chapter 1:

$$F_{SD}(E_t) = \left[1 - \frac{1}{1+(1/g_D)\exp[(E_t-E_F)/kT]}\right]$$
$$= \frac{1}{1+g_D\exp[(E_F-E_t)/(kT)]} \tag{36a}$$

for donor interface traps and

$$F_{SA}(E_t) = \frac{1}{1+g_A\exp[(E_t-E_F)/kT]} \tag{36b}$$

for acceptor interface traps, where E_t is the energy of the interface trap, and the ground-state degeneracy is 2 for donor (g_D) and 4 for acceptor (g_A). Presumably every interface has both kinds of traps. A convenient notation is to interpret the sum of these by an equivalent D_{it} distribution, with an energy level called neutral level E_0 above which the states are of acceptor type, and below which are of donor type, as shown in Fig. 13. To calculate the trapped charge, it can also be assumed that at room temper-

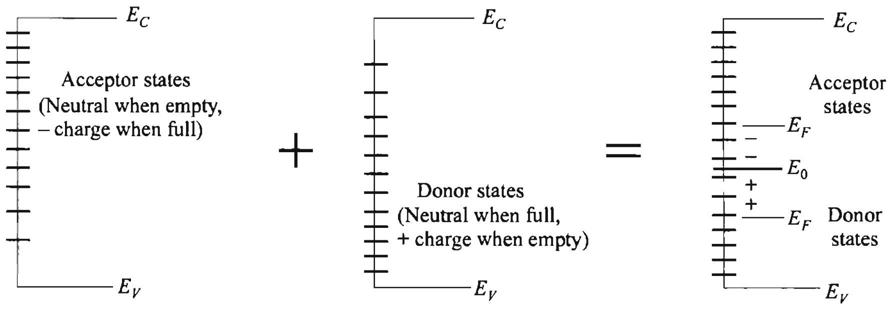


Fig. 13 Any interface-trap system consisting of both acceptor states and donor states can be interpreted by an equivalent distribution with a neutral level E_0 above which the states are of acceptor type and below which of donor type. When E_F is above (below) E_0, net charge is – (+).

ature, the occupancy takes on the value of 0 and 1 above and below E_F. With these assumptions, the interface-trapped charge can now be easily calculated by:

$$
\begin{aligned}
Q_{it} &= -q\int_{E_0}^{E_F} D_{it}dE \qquad && E_F \text{ above } E_0,\\
&= +q\int_{E_F}^{E_0} D_{it}dE \qquad && E_F \text{ below } E_0. \qquad (37)
\end{aligned}
$$

The foregoing charges are the effective net charges per unit area (i.e., C/cm^2). Because interface-trap levels are distributed across the energy bandgap, they are characterized by an interface-trap density distribution:

$$D_{it} = \frac{1}{q}\frac{dQ_{it}}{dE} \qquad \text{Number of traps/cm}^2\text{-eV}. \qquad (38)$$

This is the concept used to determine D_{it} experimentally—from the change of Q_{it} in response to the change of E_F or surface potential ψ_s. On the other hand, Eq. 38 cannot distinguish whether the interface traps are of donor type or acceptor type but only determine the magnitude of D_{it}.

When a voltage is applied, the Fermi level moves up or down with respect to the interface-trap levels and a change of charge in the interface traps occurs. This change of charge affects the MIS capacitance and alters the ideal MIS curve. The basic equivalent circuit[21] incorporating the interface-trap effect is shown in Fig. 14a. In the figure, C_i and C_D are the insulator capacitance and the semiconductor depletion-layer capacitance, respectively. C_{it} and R_{it} are the capacitance and resistance associated with the interface traps and, thus, are also functions of energy. The product $C_{it}R_{it}$ is defined as the interface-trap lifetime τ_{it}, which determines the frequency behavior of the interface traps. The parallel branch of the equivalent circuit in Fig. 14a can be converted into a frequency-dependent capacitance C_p in parallel with a frequency-dependent conductance G_p, as shown in Fig. 14b, where

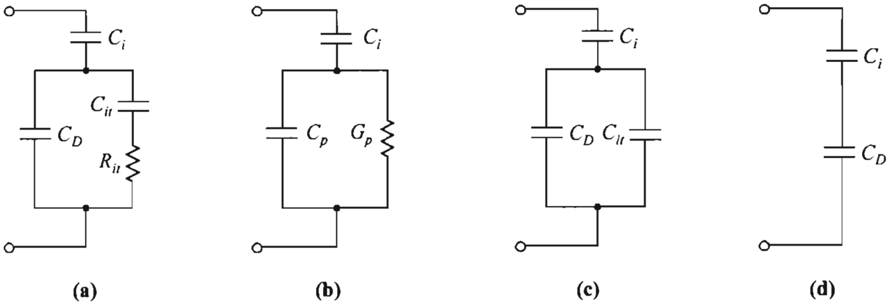


Fig. 14 (a)–(b) Equivalent circuits including interface-trap effects, C_{it} and R_{it}. (After Ref. 21.) (c) Low-frequency limit. (d) High-frequency limit.

$$C_p = C_D + \frac{C_{it}}{1 + \omega^2 \tau_{it}^2} \tag{39}$$

and

$$\frac{G_p}{\omega} = \frac{C_{it}\omega\tau_{it}}{1 + \omega^2 \tau_{it}^2} . \tag{40}$$

Also of particular interest are the equivalent circuits in the low-frequency and high-frequency limits, included in Fig. 14. In the low-frequency limit, R_{it} is set to zero and C_D is in parallel to C_{it}. In the high-frequency limit, the C_{it}-R_{it} branch is ignored or open. Physically it means that the traps are not fast enough to respond to the fast signal. The total terminal capacitance for these two cases (low-frequency C_{LF} and high-frequency C_{HF}) are;

$$C_{LF} = \frac{C_i(C_D + C_{it})}{C_i + C_D + C_{it}} , \tag{41}$$

$$C_{HF} = \frac{C_i C_D}{C_i + C_D} . \tag{42}$$

These equations and equivalent circuits will be useful in the measurement of interface traps, to be discussed next.

4.3.2 Measurement of Interface Traps

Either capacitance measurement or conductance measurement can be used to evaluate the interface-trap density, because both the input conductance and the input capacitance of the equivalent circuit contain similar information about the interface traps. It will be shown that the conductance technique can give more accurate results, especially for MOS capacitors with relatively low interface-trap density ($\approx 10^{10}$ cm^{-2}-eV^{-1}). The capacitance measurement, however, can give rapid evaluation of flat-band shift and the total interface-trapped charge.

Figure 15a shows qualitatively the high-frequency and low-frequency C-V characteristics with and without interface traps. A very noticeable effect of the interface traps is that the curves are stretched out in the voltage direction. This is due to the fact that extra charge has to fill the traps, so it takes more total charge or applied voltage to accomplish the same surface potential ψ_s (or band bending). This is demonstrated more clearly in Fig. 15b where ψ_s is plotted against the apply voltage directly, with and without interface traps. As shown later, this ψ_s-V curve can be used to determine D_{it}. Another point to be noted is the gap in capacitance between the low-frequency and high-frequency curves, before the point of $V_{\min}$ near strong inversion. This difference is proportional to D_{it}.

One other helpful point is that interface traps affect the total capacitance in two ways. A direct impact is through the extra circuit elements C_{it} and R_{it}. A second impact is indirectly on C_D. For a fixed bias, since some charge will be needed to fill the interface traps, the remaining charge to be put in the depletion layer is reduced and

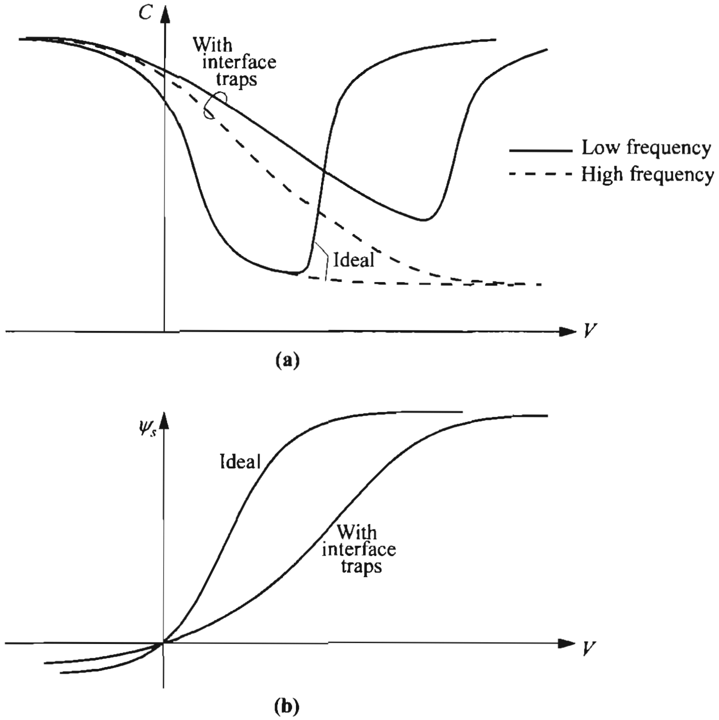


Fig. 15 (a) Influence of interface traps on high-frequency and low-frequency *C-V* curves. (b) The stretch out of *C-V* curves is due to a less effective modulation of surface potential ψ_s by the applied voltage *V*. Example is on *p*-type semiconductor.

this will reduce the surface potential ψ_s or band bending. But since the relationship between C_D and ψ_s is fixed (Eq. 22 or 28), changing ψ_s means changing C_D also. This explains that for the high-frequency limit, even though the equivalent circuit of Fig. 14d does not contain the C_{it} element, the high-frequency *C-V* curve in Fig. 15a is still affected by interface traps, through C_D.

Observing the four curves in Fig. 15a will help to understand the different capacitance methods in determining D_{it}. There are basically three methods; (1) low-frequency capacitance—to compare the measured low-frequency curve with theoretical ideal curve; (2) high-frequency capacitance—to compare the measured high-frequency curve with theoretical ideal curve; and (3) high-low–frequency capacitance—to compare the measured low-frequency to high-frequency curves.

Before we discuss each of these capacitance methods, we first derive some useful terms that are valid for all. First the relationship between C_{it} and D_{it} is derived as follows. Since $dQ_{it} = qD_{it}\,dE$, and $dE = qd\psi_s$, we obtain

$$C_{it} \equiv \frac{dQ_{it}}{d\psi_s}$$
$$= q^2 D_{it} . \tag{43}$$

Next we will derive the stretch out of the ψ_s-V curve in relationship to interface traps. Using the low-frequency equivalent circuits of Fig. 14c, the applied voltage is partitioned between the oxide layer and the semiconductor layer (Eq. 18). The portion of the voltage across the semiconductor ψ_s is simply given by the voltage divider of the capacitor network, i.e.,

$$\frac{d\psi_s}{dV} = \frac{C_i}{C_i + (C_D + C_{it})} . \tag{44}$$

Substitution of Eq. 43 into Eq. 44 gives

$$D_{it} = \frac{C_i}{q^2}\left[\left(\frac{d\psi_s}{dV}\right)^{-1} - 1\right] - \frac{C_D}{q^2} . \tag{45}$$

From this equation, D_{it} can be calculated if the ψ_s-V relationship (Fig. 15b) can be obtained from the capacitance measurement.

High-Frequency Capacitance Method. Terman[7] was the first who developed the high-frequency method. This method, as shown in the equivalent circuit of Fig. 14d, has the advantage that it does not contain the circuit element of C_{it}. The measured C_{HF}, as given by Eq. 42, can give C_D directly. Once C_D is known, ψ_s can be calculated from theory and the ψ_s-V relationship is obtained. Equation 45 is then used to determine D_{it}.

Low-Frequency Capacitance Method. Berglund[22] was the first to use integration of low-frequency capacitance to obtain the ψ_s-V relationship, which then is used for obtaining D_{it} from Eq. 45. Starting from Eq. 44 which is based on the low-frequency equivalent circuit of Fig. 14c,

$$\frac{d\psi_s}{dV} = \frac{C_i}{C_i + C_D + C_{it}} = 1 - \frac{C_D + C_{it}}{C_i + C_D + C_{it}}$$
$$= 1 - \frac{C_{LF}}{C_i} . \tag{46}$$

Integrating Eq. 46 over two applied voltages yields

$$\psi_s(V_2) - \psi_s(V_1) = \int_{V_1}^{V_2}\left(1 - \frac{C_{LF}}{C_i}\right)dV + \text{constant} . \tag{47}$$

Equation 47 indicates that the surface potential at any applied voltage can be determined by integrating the value of $(1 - C_{LF}/C_i)$. The integrand constant can be the starting point at accumulation or strong inversion where ψ_s is known and it has weak dependence on the applied voltage. Once ψ_s is known, C_D can be calculated from Eq. 45, provided that the doping profile is known. One disadvantage of the low-frequency capacitance method is the measurement difficulty in the presence of increased dc leakage for thinner oxides.

High-Low–Frequency Capacitance Method. This method combining both high-frequency and low-frequency capacitance was developed by Castagne and Vapaille.[23] The advantage of this method is that no theoretical calculation is needed for comparison, and such calculation for a nonuniform doping profile is complicated, if the profile is known at all. Starting with the equations for low-frequency and high-frequency limits (Eqs. 41 and 42), we can express

$$\begin{aligned} C_{it} &= \left(\frac{1}{C_{LF}} - \frac{1}{C_i}\right)^{-1} - C_D \\ &= \left(\frac{1}{C_{LF}} - \frac{1}{C_i}\right)^{-1} - \left(\frac{1}{C_{HF}} - \frac{1}{C_i}\right)^{-1}. \end{aligned} \tag{48}$$

Defining the capacitance gap as $\Delta C \equiv C_{LF} - C_{HF}$, and using the relationship $D_{it} = C_{it}/q^2$, we obtain the trap density directly

$$\begin{aligned} D_{it} &= \frac{C_i}{q^2}\left[\left(\frac{1}{\Delta C/C_i + C_{HF}/C_i} - 1\right)^{-1} - \left(\frac{1}{C_{HF}/C_i} - 1\right)^{-1}\right] \\ &= \frac{\Delta C}{q^2}\left(1 - \frac{C_{HF} + \Delta C}{C_i}\right)^{-1}\left(1 - \frac{C_{HF}}{C_i}\right)^{-1} \end{aligned} \tag{49}$$

for each bias point. As shown in this equation, the trap density, on the first order, is proportional to the capacitance gap ΔC. If the energy spectrum of D_{it} is to be determined, either the low-frequency capacitance integration approach or the high-frequency method can be applied to determine ψ_s.

Conductance Method. Nicollian and Goetzberger give a detailed and comprehensive discussion of the conductance method.[24] Difficulty arises in the capacitance methods because the interface-trap capacitance must be extracted from the measured capacitance which consists of oxide capacitance, depletion-layer capacitance, and interface-trap capacitance. As previously mentioned, both the capacitance and conductance as functions of voltage and frequency contain identical information about interface traps. Greater inaccuracies arise in extracting this information from the measured capacitance because the difference between two capacitances must be calculated. This difficulty does not apply to the measured conductance which is directly related to the interface traps. Thus conductance measurements yield more accurate and reliable results, particularly when D_{it} is low as in the thermally oxidized SiO_2-Si system. Figure 16 shows the measured capacitance and conductance at 5 and 100 kHz. The largest capacitance spread is only 14% while the magnitude of the conductance peak varies by over one order of magnitude in this frequency range.

The simplified equivalent circuit in Fig. 14b illustrates the principle of the MIS conductance technique. The impedance of the MIS capacitor is measured by a bridge across the capacitor terminals. The insulator capacitance C_i is also measured in the region of strong accumulation. The reactance of the insulator capacitance is subtracted from this impedance and the resulting impedance converted into an admit-

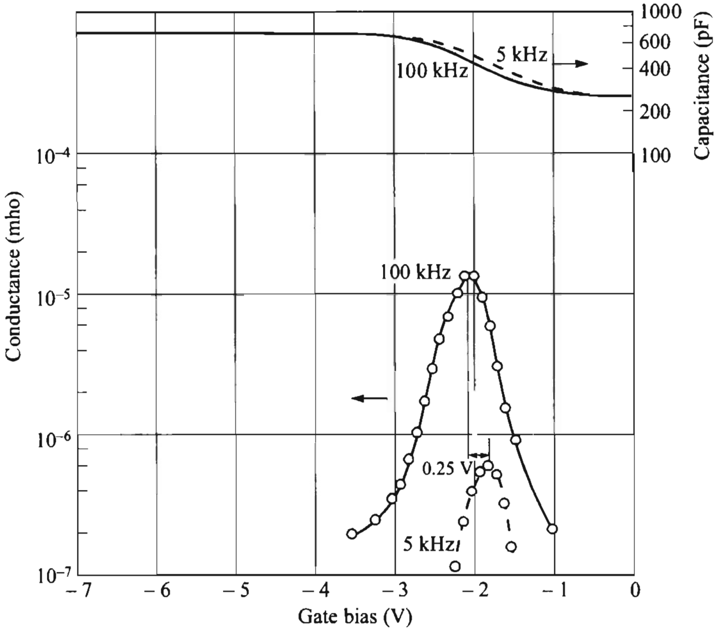


Fig. 16 Comparison of MIS capacitance and conductance measurement at two frequencies, showing that conductance is much more sensitive to frequency than to capacitance. (After Ref. 24.)

tance. This leaves C_D in parallel with the series $R_{it}C_{it}$ network of the interface traps (Fig. 14b). The equivalent parallel conductance G_p divided by ω is given by Eq. 40 which does not contain C_D and depends only on the interface-trap branch of the equivalent circuit. An expression to convert the measured admittance to the conductance of the interface-trap branch is given by

$$\frac{G_p}{\omega} = \frac{\omega C_i^2 G_{in}}{G_{in}^2 + \omega^2(C_i - C_{in})^2} = \frac{C_{it}\omega\tau_{it}}{1 + \omega^2\tau_{it}^2} \tag{50}$$

where the last term is a repeat of Eq. 40. At a given bias, G_p/ω can be measured as a function of frequency. A plot of G_p/ω versus ω goes through a maximum when $\omega\tau_{it} = 1$, and gives τ_{it} directly. The value of G_p/ω at the maximum is $C_{it}/2$. Thus, the equivalent parallel conductance corrected for C_i gives C_{it} and τ_{it} ($= R_{it}C_{it}$) directly from the measured conductance. Once C_{it} is known, the interface-trap density is obtained by using the relation $D_{it} = C_{it}/q^2$.

Typical results in a Si-SiO$_2$ system[25] show that near the midgap D_{it} is relatively constant, but it increases toward the conduction- and valence-band edges. Orientation dependence is particularly important. In (100) orientation D_{it} is about an order of magnitude smaller than that in (111). This result has been correlated with the available bonds per unit area on the silicon surface.[26,27] Table 1 shows the properties of

Table 1 Properties of Silicon Crystal Planes

Orientation	Plane area of unit cell	Atoms in cell area	Available bonds in cell area	Atoms/cm^2	Available bonds/cm^2
$\langle 111 \rangle$	$\sqrt{3}a^2/2$	2	3	7.85×10^{14}	11.8×10^{14}
$\langle 110 \rangle$	$\sqrt{2}a^2$	4	4	9.6×10^{14}	9.6×10^{14}
$\langle 100 \rangle$	a^2	2	2	6.8×10^{14}	6.8×10^{14}

silicon crystal planes oriented along (111), (110), and (100) directions. It is apparent that the (111) surface has the largest number of available bonds per area, and the (100) surface has the smallest. One would also expect that the (100) surface has the lowest oxidation rate which is advantageous for thin oxides. If we assume that the origin of interface traps is due to excess silicon in the oxide, then the lower the oxidation rate the smaller the amount of the excess silicon; thus the (100) surface should have the smallest interface-trap density. Therefore, all modern silicon MOSFETs are fabricated on (100)-oriented substrates.

Interface traps in the Si-SiO$_2$ system comprise of many levels. These are so closely spaced in energy that they cannot be distinguished as separate levels and actually appear as a continuum over the bandgap of the semiconductor. The equivalent circuit for an MIS capacitor with a single-level time constant (Fig. 14a) should, therefore, be interpreted as for a certain bias or trap level.

Figure 17 shows the variation of the time constant τ_{it} versus surface potential (or trap level) for MOS capacitors with steam-grown oxides on (100) silicon substrates, where $\bar{\psi}_s$ is the average surface potential (to be discussed later). These curves can be fitted by the following expressions:

$$\tau_{it} = \frac{1}{\bar{v}\sigma_p n_i}\exp\left[-\frac{q(\psi_{Bp}-\bar{\psi}_s)}{kT}\right] \quad \text{for } p\text{-type} \tag{51a}$$

$$\tau_{it} = \frac{1}{\bar{v}\sigma_n n_i}\exp\left[-\frac{q(\psi_{Bn}+\bar{\psi}_s)}{kT}\right] \quad \text{for } n\text{-type} \tag{51b}$$

where σ_p and σ_n are the capture cross sections of holes and electrons respectively, and $\bar{v}$ is the average thermal velocity. These results indicate that the capture cross section is independent of energy. The capture cross sections obtained[24] from Fig. 17 are $\sigma_p = 4.3\times10^{-16}$ cm^2 and $\sigma_n = 8.1\times10^{-16}$ cm^2, where the value of $\bar{v} = 10^7$ cm/s has been used. For (111)-oriented silicon the variation of time constant versus surface potential is similar to that of (100) and the measured capture cross sections are smaller with $\sigma_p = 2.2\times10^{-16}$ cm^2 and $\sigma_n = 5.9\times10^{-16}$ cm^2.

We must also consider the statistical fluctuation of surface potential due to surface charges which include the fixed oxide charges Q_f and the interface-trapped charges Q_{it}. From Eq. 51b, a small fluctuation in $\bar{\psi}_s$ causes a large fluctuation in τ_{it}. Assuming that surface charges are randomly distributed in the plane of the interface, the electric field at the semiconductor surface will fluctuate over the plane of the

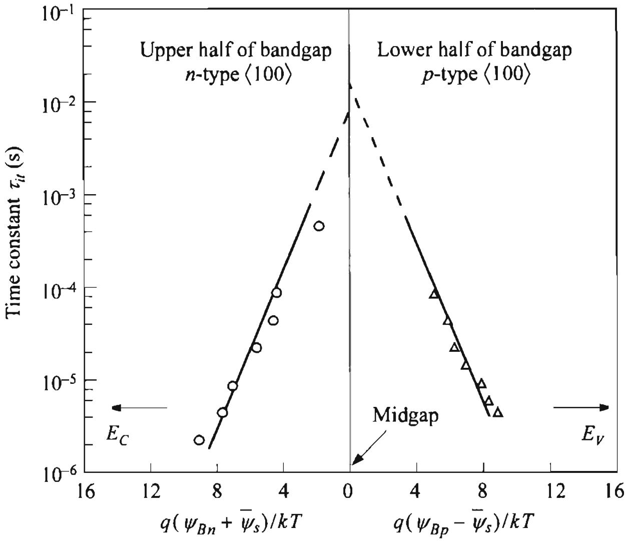


Fig. 17 Variation of trap time constant τ_{it} vs. energy. $T = 300$ K. (After Ref. 24.)

interface. Figure 18 shows calculated values of G_p/ω as a function of frequency for a Si-SiO$_2$ MOS capacitor biased in depletion and weak inversion, including the effect of time-constant dispersion resulting from the interface-trap continuum and the statistical (Poisson) distribution of surface charges (Q_{it} + Q_f). Experimental results are also

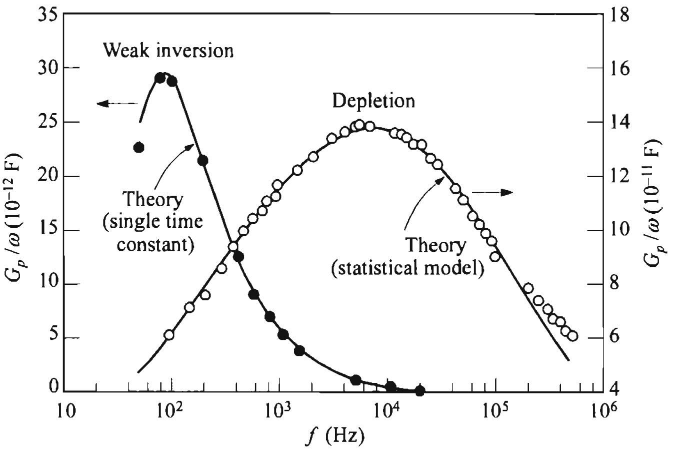


Fig. 18 G_p/ω vs. frequency for a Si-SiO$_2$ MOS capacitor biased in depletion region (broad curve) and weak-inversion region (narrow curve). Circles are experimental results. Lines are theoretical calculations. (After Ref. 24.)

shown (open and solid circles); their excellent agreement with the statistical results indicates the importance of the statistical model.

The impact of charge or potential fluctuation is also implicated in Fig. 18. In depletion, the potential fluctuation broadens the frequency range but the peak frequency is unaffected, which is the most-important parameter to be extracted. On the other hand, in weak inversion, potential fluctuation has a more profound effect. This is because potential fluctuation will cause some local regions to be in depletion and conductance in those regions dominates disproportionally. So even though the G_p/ω curve is not broadened and can be characterized by a single time constant, its value is shifted by an amount depending on the statistical nature of the charge fluctuation. To circumvent this problem, both *n*- and *p*-type devices can be used in the depletion-region measurement only to get the trap spectrum over the two halves of the bandgap.

4.3.3 Oxide Charges and Work-Function Difference

Oxide charges, other than that of the interface traps, include the fixed oxide charge Q_f, the mobile ionic charge Q_m, and the oxide trapped charge Q_{ot}, as shown in Fig. 12. These will be discussed in sequence. In general, unlike interface-trapped charges, these oxide charges are independent of bias, so they cause a parallel shift in the gate-bias direction, as indicated in Fig. 19a. The flat-band voltage shift due to any oxide charge is given by Gauss' law;

$$\Delta V = -\frac{1}{C_i}\left[\frac{1}{d}\int_0^d x\rho(x)dx\right] \tag{52}$$

where $\rho(x)$ is the charge density per unit volume. The effect on the voltage shift is weighted according to the location of the charge, i.e., the closer to the oxide-semiconductor interface, the more shift it will cause. Qualitatively the influence of positive oxide charges can be explained in Figs. 19b–d. Positive charge is equivalent to an added positive gate bias for the semiconductor so it requires a more negative gate bias to achieve the same original semiconductor band bending. Notice that in the new flat-band condition (Fig. 19d), the oxide field is no longer zero.

The fixed oxide charge Q_f has the following properties: It is located very close to the Si-SiO$_2$ interface;[9] it is generally positive; its density is not greatly affected by the oxide thickness or by the type or concentration of impurities in the silicon, but it depends on oxidation and annealing conditions, and on silicon surface orientation. It has been suggested that excess silicon (trivalent silicon) or the loss of an electron from excess oxygen centers (nonbridging oxygen) near the Si-SiO$_2$ interface is the origin of fixed oxide charge. In electrical measurements, Q_f can be regarded as a charge sheet located at the Si-SiO$_2$ interface,

$$\Delta V_f = -\frac{Q_f}{C_i}. \tag{53}$$

Mobile ionic charges can move back and forth through the oxide layer, depending on biasing conditions, and thus give rise to voltage shifts. The shift usually is

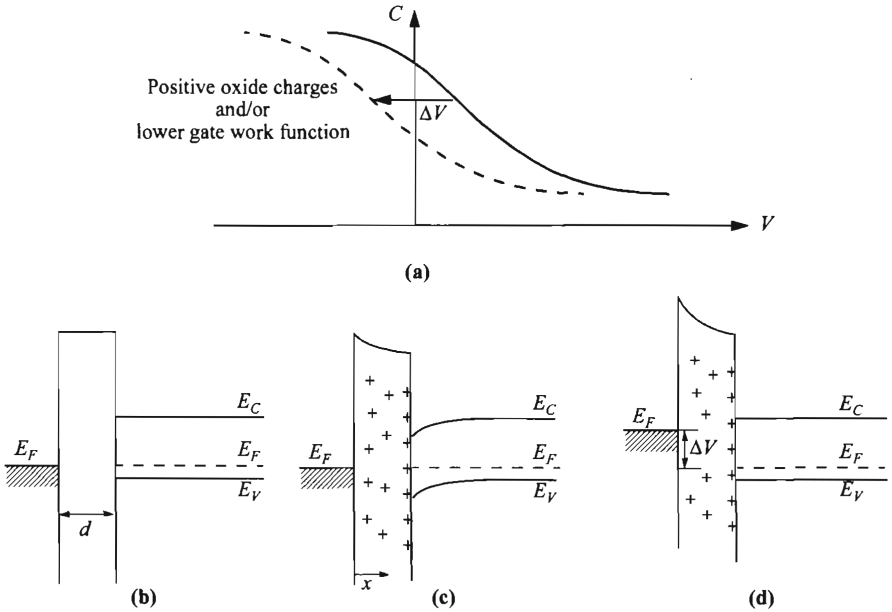


Fig. 19 (a) High-frequency C-V curve (on p-semiconductor), shifted along the voltage axis due to positive oxide charges. (b) Band diagram at flat band, original. (c) With positive oxide charges and (d) new flat-band bias.

enhanced at elevated temperature. In severe cases, hysteresis can be seen when the gate voltage is swept in opposite polarities. It was first demonstrated by Snow et al.[28] that alkali ions, such as sodium, in thermally grown SiO_2 films are mainly responsible for the instability of the oxide-passivated devices. Reliability problems in semiconductor devices operated at high temperatures and voltages may be related to trace contamination by alkali metal ions. The voltage shift is given by Eq. 52 and is represented by

$$\Delta V_m = -\frac{Q_m}{C_i} \tag{54}$$

where Q_m is the effective net charge of mobile ions per unit area at the Si-SiO_2 interface and the actual mobile ions $\rho(x)$ is used.

To prevent mobile ionic charge contamination of the oxide during device life, one can protect it with a film impervious to mobile ions such as amorphous or small-crystallite silicon nitride. For amorphous Si_3N_4, there is very little sodium penetration. Other sodium barrier layers include Al_2O_3 and phosphosilicate glass.

Oxide trapped charge is associated with defects in SiO_2. The oxide traps are usually initially neutral and are charged by introducing electrons and holes into the oxide layer. This can occur from any current passing through the oxide layer (to be discussed in next section), hot-carrier injection, or by photon excitation. The shift due to the oxide trapped charge is again given by Eq. 52,

$$\Delta V_{ot} = -\frac{Q_{ot}}{C_i} \tag{55}$$

where Q_{ot} is the effective net charge per unit area at the Si-SiO$_2$ interface.

The total voltage shift due to all the oxide charges is the sum

$$\Delta V = \Delta V_f + \Delta V_m + \Delta V_{ot} = -\frac{Q_f + Q_m + Q_{ot}}{C_i}. \tag{56}$$

Work-Function Difference. For the preceding discussions on ideal MIS capacitor, it has been assumed that the work-function difference for a p-type semiconductor (Fig. 2b)

$$\phi_{ms} \equiv \phi_m - \left(\chi + \frac{E_g}{2q} + \psi_{Bp}\right) \tag{57}$$

is zero. If the value of ϕ_{ms} is not zero, the experimental C-V curve will be shifted from the theoretical curve by the same amount in gate bias, as indicated in Fig. 20. This shift is in addition to the oxide charges, so the net flat-band voltage becomes

$$V_{FB} = \phi_{ms} - \frac{Q_f + Q_m + Q_{ot}}{C_i}. \tag{58}$$

Figure 21 demonstrates the correlation of flat-band voltage with metal work function determined by different means. The energy band for the Si-SiO$_2$ interface has been obtained from electron photoemission measurements;[30] SiO$_2$ bandgap is found to be about 9 eV, and the electron affinity ($q\chi_i$) is 0.9 eV. From photoresponse versus photon energy on various metals,[29] the intercept on the $h\nu$ axis corresponds to the metal-SiO$_2$ barrier energy $q\phi_B$. The metal work function is given by the sum of ϕ_B and χ_i (refer to Fig. 2). The metal work functions as obtained from the photoresponse and the capacitance curves are in excellent agreement.

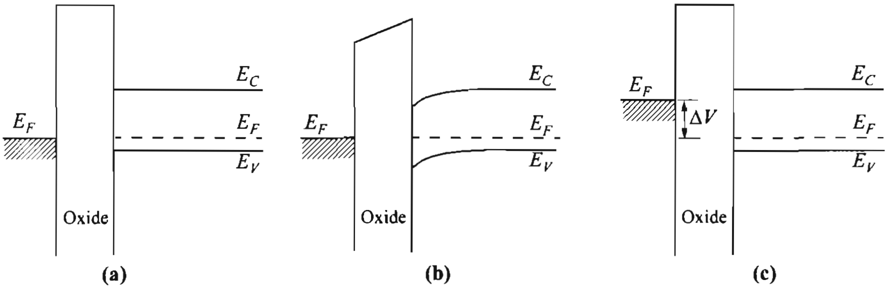


Fig. 20 (a) Band diagram at flat band, $\phi_{ms} = 0$. (b) With a lower gate work function, zero bias, and (c) new flat-band bias.

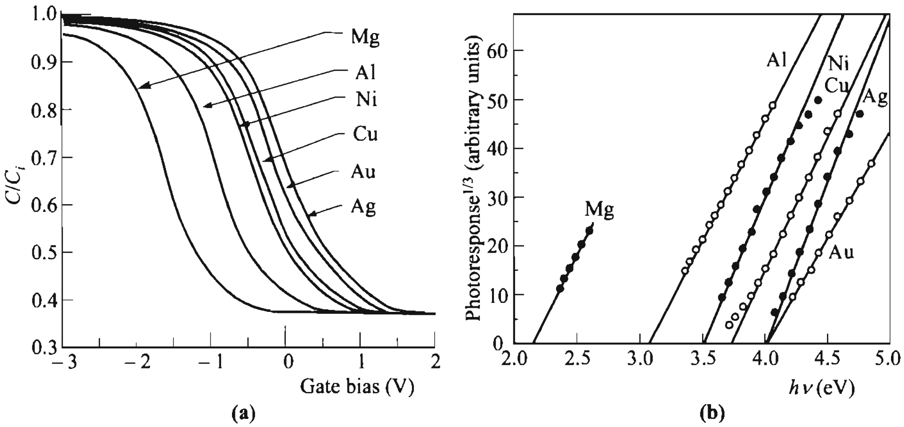


Fig. 21 Correlation of (a) flat-band voltage from capacitance measurement and (b) barrier height from photoresponse. (After Ref. 29.)

In modern integrated-circuit processing, heavily doped polysilicon has been used to replace Al as the gate electrode. For an n^+-polysilicon gate, the Fermi level essentially coincides with the bottom of the conduction band E_C and the effective work function ϕ_m is equal to the Si electron affinity (χ_{Si} = 4.05 V). For a p^+-polysilicon gate, the Fermi level coincides with the top of the valence band E_V and the effective work function ϕ_m is equal to the sum of χ_{Si} and E_g/q (5.17 V). This is one of the advantages of using poly-Si gates in MOSFETs since the same material can give different work functions by doping. Figure 22 shows the work-function difference as a

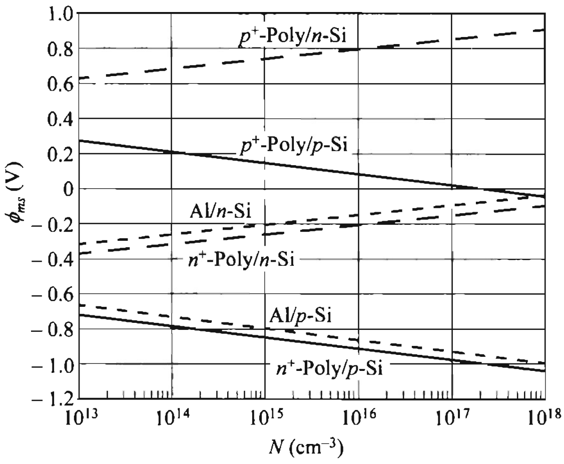


Fig. 22 Work-function difference ϕ_{ms} vs. doping, for gate electrodes of degenerate polysilicon and Al on p- and n-Si.

function of Si doping concentration for Al, Au, p^+-, and n^+-polysilicon gates. By an appropriate choice of gate electrode, both n- and p-type silicon surfaces can be varied from accumulation to inversion.

4.3.4 Carrier Transport

In an ideal MIS capacitor the conductance of the insulating film is assumed to be zero. Real insulators, however, show some degree of carrier conduction when the electric field or temperature is sufficiently high. To estimate the electric field in an insulator under biasing conditions, we obtain

$$\mathscr{E}_i = \mathscr{E}_s\left(\frac{\varepsilon_s}{\varepsilon_i}\right) \approx \frac{V}{d} \tag{59}$$

where $\mathscr{E}_i$ and $\mathscr{E}_s$ are the electric fields in the insulator and the semiconductor respectively, and ε_i and ε_s are the corresponding permittivities. The equation also assumes negligible oxide charges and that the flat-band voltage and the semiconductor band bending ψ_s are small compared to the applied voltage. Table 2 summarizes the basic conduction processes in insulators. It also emphasizes the voltage and temperature dependence of each process that are used often to identify the exact conduction mechanism experimentally.

Tunneling is the most-common conduction mechanism through insulators under high fields. The tunnel emission is a result of quantum mechanics by which the elec-

Table 2 Basic Conduction Processes in Insulators

Process	Expression	Voltage & temperature dependence
Tunneling	$J \propto \mathscr{E}_i^2 \exp\left[-\frac{4\sqrt{2m^*}(q\phi_B)^{3/2}}{3q\hbar\mathscr{E}_i}\right]$	$\propto V^2 \exp\left(\frac{-b}{V}\right)$
Thermionic emission	$J = A^{**}T^2 \exp\left[\frac{-q(\phi_B - \sqrt{q\mathscr{E}_i/4\pi\varepsilon_i})}{kT}\right]$	$\propto T^2 \exp\left[\frac{q}{kT}(a\sqrt{V} - \phi_B)\right]$
Frenkel-Poole emission	$J \propto \mathscr{E}_i \exp\left[\frac{-q(\phi_B - \sqrt{q\mathscr{E}_i/\pi\varepsilon_i})}{kT}\right]$	$\propto V \exp\left[\frac{q}{kT}(2a\sqrt{V} - \phi_B)\right]$
Ohmic	$J \propto \mathscr{E}_i \exp\left(\frac{-\Delta E_{ac}}{kT}\right)$	$\propto V \exp\left(\frac{-c}{T}\right)$
Ionic conduction	$J \propto \frac{\mathscr{E}_i}{T} \exp\left(\frac{-\Delta E_{ai}}{kT}\right)$	$\propto \frac{V}{T} \exp\left(\frac{-d'}{T}\right)$
Space-charge-limited	$J = \frac{9\varepsilon_i \mu V^2}{8d^3}$	$\propto V^2$

A^{**} = effective Richardson constant. ϕ_B = barrier height. $\mathscr{E}_i$ = electric field in insulator. ε_i = insulator permittivity. $m*$ = effective mass. d = insulator thickens. ΔE_{ac} = activation energy of electrons. ΔE_{ai} = activation energy of ions. $V \approx \mathscr{E}_i d$. $a \equiv \sqrt{q/4\pi\varepsilon_i d}$. b, c, and d' are constants.

tron wave function can penetrate through a potential barrier (see Section 1.5.7). It has the strongest dependence on the applied voltage but is essentially independent of the temperature. According to Fig. 23 tunneling can be divided into direct tunneling and Fowler-Nordheim tunneling where carriers tunnel through only a partial width of the barrier.[31]

The Schottky emission process is similar to the process discussed in Chapter 3, where thermionic emission over the metal-insulator barrier or the insulator-semiconductor barrier is responsible for carrier transport. In Table 2, the term subtracting from ϕ_B is due to image-force lowering (see Section 3.2.4). A plot of $\ln(J/T^2)$ versus $1/T$ yields a straight line with a slope determined by the net barrier height.

The Frenkel-Poole emission,[32,33] shown in Fig. 23d, is due to emission of trapped electrons into the conduction band. The supply of electrons from the traps is through thermal excitation. For trap states with Coulomb potentials, the expression is similar to that of the Schottky emission. The barrier height, however, is the depth of the trap potential well. The barrier reduction is larger than in the case of Schottky emission by a factor of 2, since the barrier lowering is twice as large due to the immobility of the positive charge.

At low voltage and high temperature, current is carried by thermally excited electrons hopping from one isolated state to the next. This mechanism yields an ohmic characteristic exponentially dependent on temperature.

The ionic conduction is similar to a diffusion process. Generally, the dc ionic conductivity decreases during the time the electric field is applied because ions cannot be readily injected into or extracted from the insulator. After an initial current flow, positive and negative space charges will build up near the metal-insulator and the semiconductor-insulator interfaces, causing a distortion of the potential distribution. When the applied field is removed, large internal fields remain which cause some, but

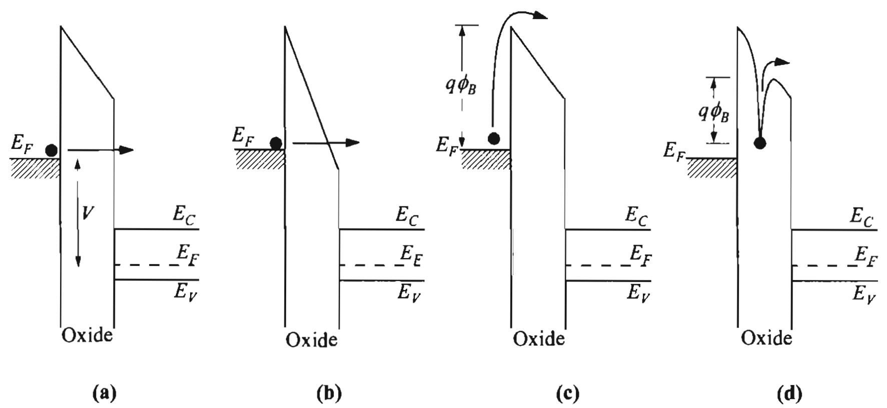


Fig. 23 Energy-band diagrams showing conduction mechanisms of (a) direct tunneling, (b) Fowler-Nordheim tunneling, (c) thermionic emission, and (d) Frenkel-Poole emission.

not all, ions to flow back toward their equilibrium position. Because of this, hysteresis results in I-V traces.

The space-charge-limited current results from carriers injected into a lightly doped semiconductor or an insulator, where no compensating charge is present. The current for the unipolar trap-free case is proportional to the square of the applied voltage. Notice that the mobility regime is relevant here (see Section 1.5.8) since mobility is typically very low in insulators.

For ultra-thin insulators, tunneling increases such that the conduction approaches that of the metal-semiconductor contact (see Section 3.3.6) where the barrier is measured at the semiconductor surface instead of the insulator and the thermionic-emission current is multiplied by a tunneling factor.

For a given insulator, each conduction process may dominate in certain temperature and voltage range. The processes are also not exactly independent of one another and should be carefully examined. For example, for the large space-charge effect, the tunneling characteristic is found to be very similar to the Schottky-type emission.[34] Figure 24 shows plots of current density versus $1/T$ for three different insulators, Si_3N_4, Al_2O_3, and SiO_2. The conduction here can generally be divided into three temperature range. At high temperatures (and high fields), the current J_1 is due to Frenkel-Poole emission. At low temperatures, the conduction is tunneling limited (J_2) which is temperature insensitive. One can also observe that the tunneling current strongly depends on the barrier height, which is related to the energy gap of the insulators. At intermediate temperatures, the current J_3 is ohmic in nature.

An example showing different conduction processes at different bias is shown in Fig. 25. Note that the two curves of opposite polarities are virtually identical. The slight difference (especially at low fields) is believed to be mainly due to the difference in barrier heights at the gold-nitride and nitride-silicon interfaces. In high electric fields the current varies exponentially with the square root of the field, a characteristic of Frenkel-Poole emission. At low fields, the characteristic is ohmic. It has been found that at room temperature for a given field, the characteristics of current density versus field are essentially independent of the film thickness, electrode materials, and polarity of the electrodes. These results strongly suggest that the current is bulk-controlled rather than electrode-controlled as in Schottky-barrier diodes.

4.3.5 Nonequilibrium and Avalanche

Going back to the capacitance curve-(d) of Fig. 7, we have a nonequilibrium condition such that the depletion width is larger than the maximum value W_{Dm} at equilibrium. This condition is called deep depletion. As the bias is swept from depletion to strong inversion, a large concentration of minority carriers is needed at the semiconductor surface. This supply of minority carriers is limited by the thermal generation rate. For a fast sweep rate, the thermal generation rate cannot keep up with the demand and deep depletion occurs. This phenomenon can also be explained by the charge placement shown in Fig. 8. The energy-band diagram for deep depletion is shown in Fig. 26a. Equilibrium condition (Fig. 26b) can be restored by slowing or

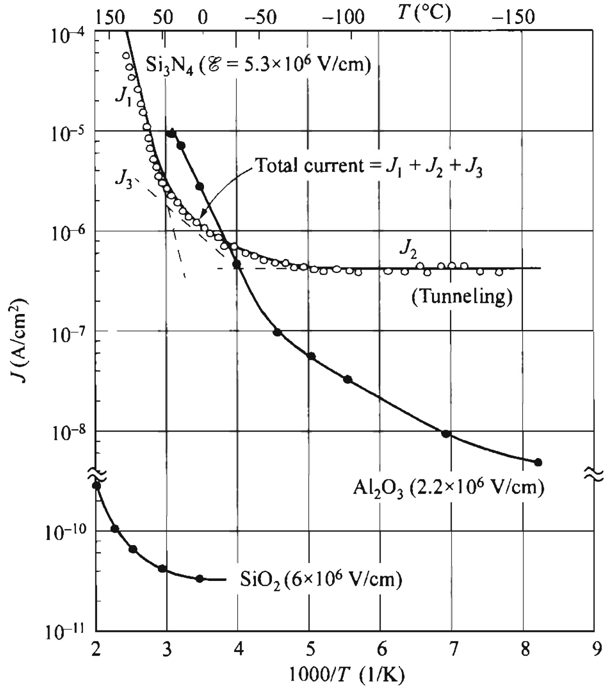


Fig. 24 Current density vs. $1/T$ for Si_3N_4, Al_2O_3, and SiO_2 films. (After Refs. 35–37.)

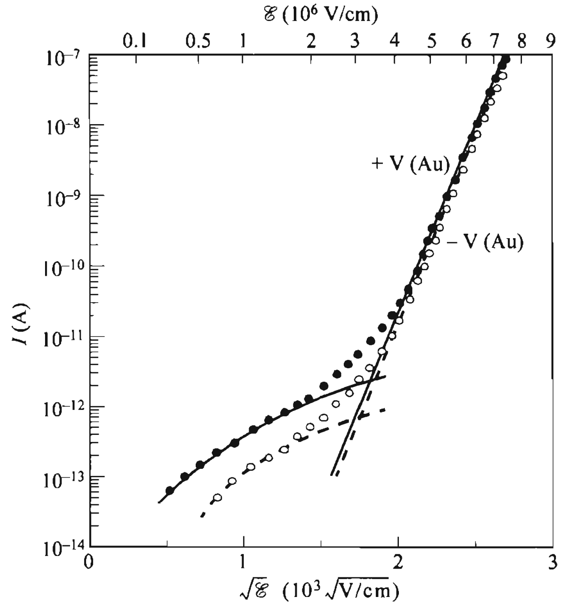


Fig. 25 Current-voltage characteristics of Au-Si_3N_4-Si capacitor at room temperature. (After Ref. 35.)

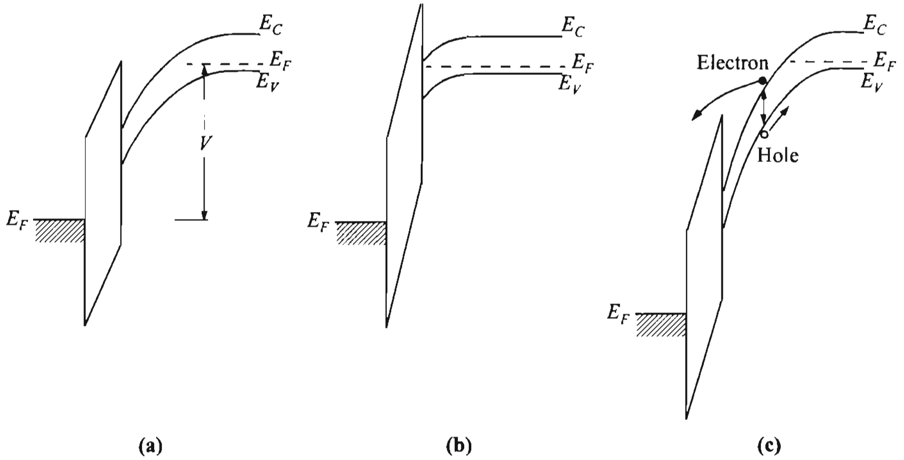


Fig. 26 Energy-band diagrams for MOS capacitor in (a) deep depletion (nonequilibrium), (b) equilibrium, and (c) deep depletion and avalanche injection of electrons into the oxide at higher bias.

stopping the voltage ramp, by raising the temperature for larger thermal generation rate, or by shining light to produce additional electron-hole pairs. When switched to equilibrium, the field is redistributed, most of which is across the oxide layer.

If driven into deep depletion with a sufficiently large bias, avalanche multiplication and breakdown can occur in the semiconductor side (Fig. 26c), similar to that in a *p-n* junction. The breakdown voltage is defined as the gate voltage that makes the ionization integral equal to unity, when integrated along a path from the semiconductor surface to the depletion-layer boundary. The avalanche breakdown voltage in the MOS capacitor under the deep-depletion condition has been calculated based on a two-dimensional model.[38] The results are shown in Fig. 27 for different doping levels and oxide thickness. It is interesting to compare these breakdown voltages to those of *p-n* junctions in Fig. 16a of Chapter 2. Bear in mind that for similar fields within the semiconductor, an MOS structure takes a higher bias because of the additional voltage taken up in the oxide layer. Several interesting features in Fig. 27 should be pointed out. First, the breakdown voltage V_{BD}, as a function of doping level, has a valley before it goes up again. The decrease of V_{BD} is the same trend as in a *p-n* junction due to the increased field with doping. The rise after the minimum is because at high doping levels, the higher field at the semiconductor surface at breakdown induces a larger voltage across the oxide layer, leading to a higher terminal voltage. Another point is that for lower impurity concentrations, the MOS breakdown is actually smaller than those of *p-n* junctions. This is due to the inclusion of the edge effect in this study. Near the perimeter of the gate electrode, the field is higher due to the two-dimensional effect which leads to a lower breakdown voltage.

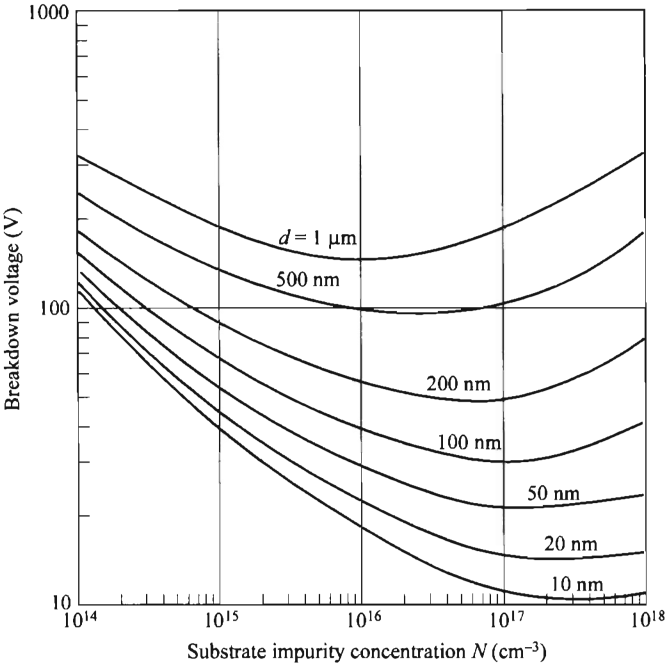


Fig. 27 Breakdown voltage of MOS capacitor in deep-depletion condition vs. silicon doping concentration, with oxide thickness as the parameter. Edge effect causing lower breakdown has been included. (After Ref. 38.)

Because of avalanche multiplication, reliability also becomes an issue due to the injection of carriers,[39] shown in Fig. 26c. Carriers generated by avalanche multiplication in the surface depletion layer, electrons in this example, will have enough energy to surmount the interfacial energy barrier and enter into the oxide layer. The energy barrier for electron injection is 3.2 eV (i.e., $q\chi_{Si} - q\chi_i = 4.1 - 0.9$), whereas for hole injection (on *n*-type substrate) it is 4.7 eV $\{$i.e., $[E_g(SiO_2) + q\chi_i] - [E_g(Si) + q\chi_{Si}]\}$. So electrons have a higher injection probability because of the lower energy barrier. The passage of hot electrons into the oxide layer generally create fixed charge, bulk and interface traps in the oxide.[9]

Hot-carrier or avalanche injection is closely related to many MOS device operations. For example, in a MOSFET, channel carriers can be accelerated by the source-to-drain electric field to have sufficient energy to surmount the Si-SiO$_2$ interfacial energy barrier. These effects are undesirable because they create a change in device characteristics during operation. On the other hand, these phenomena can be utilized in nonvolatile semiconductor memories (see Section 6.7).

Another source of hot carriers is ionization radiation such as X-ray[40] or γ-ray.[41] Ionization radiation creates electron-hole pairs in the oxide by breaking Si-O bonds.

The electric field applied across the oxide during radiation exposure drives the generated carriers in opposite directions. The electrons are considerably more mobile than the holes as they rapidly drift toward the positive electrode where most flow out into the external circuit; the holes drift much more slowly toward the negative electrode and some become trapped. The trapped holes constitute the radiation-induced positive oxide charge often observed. These trapped holes may also be responsible for the increased interface-trap density usually associated with ionizing radiation.[9]

Under optical illumination, the main effect on the MIS capacitance curves is that the capacitance in the strong-inversion region approaches the low-frequency value as the intensity of illumination is increased. Two basic mechanisms are responsible for this effect. The first is the decrease in the time constant of minority-carrier generation in the inversion layer.[12] The second is the generation of electron-hole pairs by photons, which causes a decrease of the surface potential ψ_s under constant applied voltage. This decrease of ψ_s results in a reduction of the depletion width with a corresponding increase of the capacitance. The second mechanism is dominant when the measurement frequency is high. Also under the condition of deep depletion caused by a fast gate sweep [curve-(d) in Fig. 7], the extra electron-holes pairs can supply carriers for maintaining equilibrium and curve-(d) will collapse to curve-(c).

4.3.6 Accumulation- and Inversion-Layer Thickness

For an MIS capacitor, the maximum capacitance is equal to ε_i/d which implies that charges on both sides of the electrodes cling to the two interfaces of the insulator. While such an assumption is valid on the metal-insulator interface, detailed examination on the insulator-semiconductor interface reveals that it can lead to considerable error, especially for thin oxides. This is due to charges on the semiconductor side, either accumulation or strong-inversion charges, have a distribution as a function of distance from the interface. Effectively this would reduce the maximum capacitance given by ε_i/d. For the sake of simplicity, we will discuss accumulation in the following section, but the result could also be applied to the strong-inversion case.

Classical Model. The charge distribution is controlled by the Poisson equation. Using Boltzmann statistics,

$$p(x) = N_A \exp\left(-\frac{q\psi_p}{kT}\right) \tag{60}$$

(for accumulation ψ_p is negative), the Poisson equation becomes

$$\frac{d^2\psi_p}{dx^2} = -\frac{\rho}{\varepsilon_s} \approx -\frac{qN_A}{\varepsilon_s}\exp\left(-\frac{q\psi_p}{kT}\right). \tag{61}$$

The solution of the above equation is[42]

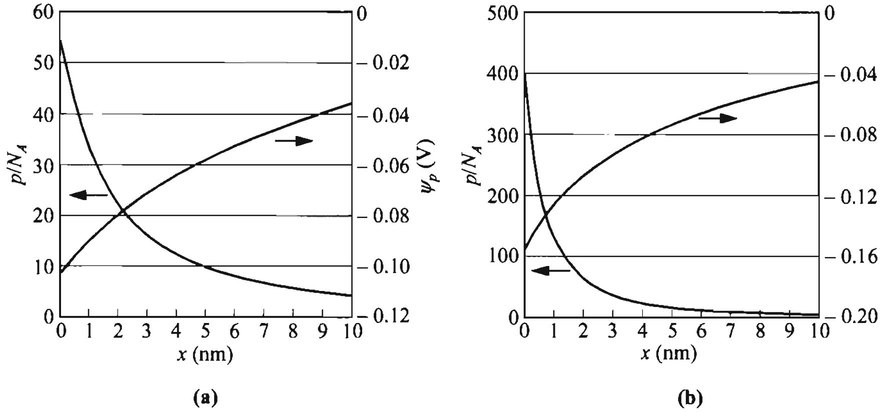
$$\psi_p(x) = -\frac{kT}{q}\ln\left(\sec^2\left\{\cos^{-1}\left[\exp\left(\frac{q\psi_p}{2kT}\right)\right] - \frac{x}{\sqrt{2}L_D}\right\}\right). \tag{62}$$

The total accumulation layer thickness where ψ_p approaches zero is equal to $\pi L_D / \sqrt{2}$ which is in the order of a few tens of nm. However, most of the carriers are confined very close to the surface. Figure 28 shows the potential and carrier distributions for two different biases. It shows that although the concentration peaks at the surface, it spreads out with an effective distance of the order of a few nm. This spread is also a function of the bias; higher bias forces the carriers to be closer to the interface.

Quantum-Mechanical Model. In quantum mechanics, the wavefunction associated with the carriers is near zero at the insulator-semiconductor interface because of the high barrier of the insulator. As a consequence, the carrier concentration peaks at some finite distance from the interface. This distance is approximately 10 Å. Figure 29 shows the results of the quantum-mechanical calculation. Macroscopically, this effect can be interpreted as a degradation in oxide capacitance (or thicker oxide). Ten Å of Si is equivalent to 3 Å of SiO_2, taking into account the difference in dielectric constant. This amount adds to the oxide thickness and lowers the capacitance. Also shown in the figure is the classical calculation. The quantum effect is shown to cause more pronounced degradation than the classical model. Another factor that causes further reduction of the capacitance is the polysilicon gates widely used in commercial technologies. Even if the polysilicon is degenerately doped, the depletion-layer and accumulation-layer thicknesses are still finite.

4.3.7 Dielectric Breakdown

One common concern for an MOS device is reliability.[44,45] Under a large bias, some current will conduct through the insulator, most commonly a tunneling current. These energetic carriers cause defects in the bulk of the dielectric film. When these defects reach a critical density level, catastrophic breakdown occurs. Microscopically, a per-



(a) (b)

Fig. 28 Classical calculation of potential and carrier profiles, with a surface potential ψ_s of (a) $4kT/q$ and (b) $6kT/q$.

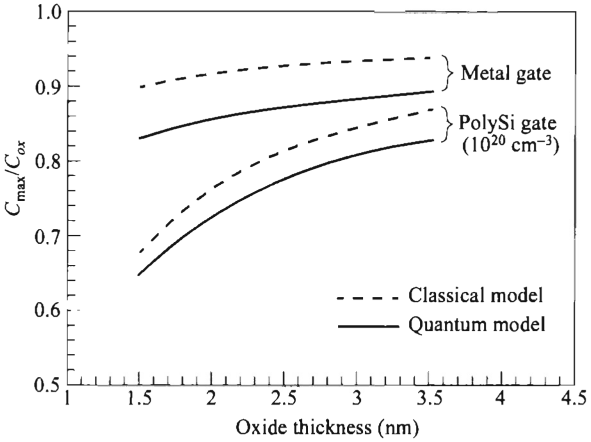


Fig. 29 Quantum-mechanical calculation of capacitance reduction. Also shown are results from classical model and those including depletion effect from polysilicon gate. (After Ref. 43.)

colation theory is used to explain breakdown (Fig. 30). On the passage of energetic carriers, defects are generated randomly. When defects are dense enough to form a continuous chain connecting the gate to the semiconductor, a conduction path is created and catastrophic breakdown occurs.

A measure to quantify reliability is time to breakdown, t_{BD}, which is the total stress time until breakdown occurs. An alternate quantity is called charge to breakdown q_{BD}, which is the total charge (integrating the current) passed through the device within t_{BD}. Obviously t_{BD} and q_{BD} are both function of applied bias. An example for t_{BD} verses oxide field for different oxide thickness is shown in Fig. 31. The plots of q_{BD} would show similar shapes and trend. A few key points can be noticed in this figure. First, t_{BD} is a function of bias. Even for a small bias, eventually the oxide will break down, taking a very long time. Conversely, a large field can be sustained for a very short time without breaking down. To search for the breakdown

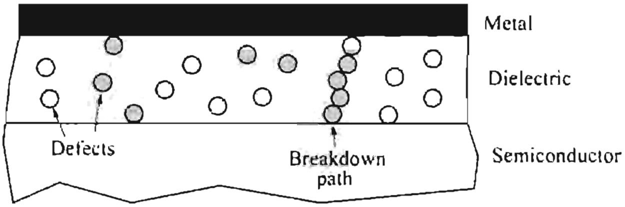


Fig. 30 Percolation theory: breakdown occurs when random defects form a chain between the gate and the semiconductor.

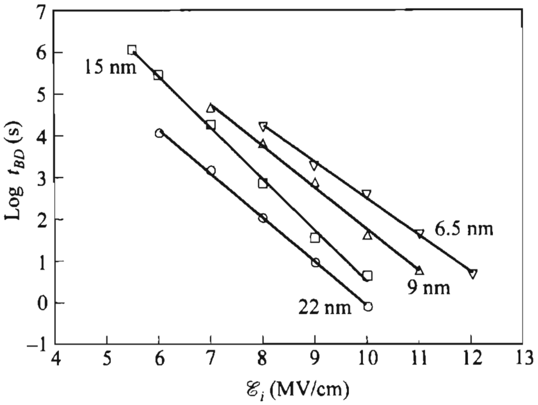


Fig. 31 Time to breakdown t_{BD} vs. oxide field, for different oxide thickness. (After Ref. 46.)

field quickly, typically a voltage ramp is applied until a large current is detected. For a common measurement, the ramping rate is typically in the order of 1 V/s. The figure shows that for this time frame, the breakdown field is around 10 MV/cm. As the oxide thickness becomes thinner, the breakdown field increases, as indicated in Fig. 31. More-recent results, however, show that this breakdown field would drop for thicknesses below ≈ 4 nm, due to an increase of tunneling current.[44]

REFERENCES

1. J. L. Moll, "Variable Capacitance with Large Capacity Change," *Wescon Conv. Rec.*, Pt. 3, p. 32 (1959).
2. W. G. Pfann and C. G. B. Garrett, "Semiconductor Varactor Using Space-Charge Layers," *Proc. IRE*, **47**, 2011 (1959).
3. D. R. Frankl, "Some Effects of Material Parameters on the Design of Surface Space-Charge Varactors," *Solid-State Electron.*, **2**, 71 (1961).
4. R. Lindner, "Semiconductor Surface Varactor," *Bell Syst. Tech. J.*, **41**, 803 (1962).
5. J. R. Ligenza and W. G. Spitzer, "The Mechanisms for Silicon Oxidation in Steam and Oxygen," *J. Phys. Chem. Solids*, **14**, 131 (1960).
6. D. Kahng and M. M. Atalla, "Silicon-Silicon Dioxide Field Induced Surface Devices," *IRE-AIEE Solid-State Device Res. Conf.*, Carnegie Inst. of Technology, Pittsburgh, PA, 1960.
7. L. M. Terman, "An Investigation of Surface States at a Silicon/Silicon Dioxide Interface Employing Metal-Oxide-Silicon Diodes," *Solid-State Electron.*, **5**, 285 (1962).
8. K. Lehovec and A. Slobodskoy, "Field-Effect Capacitance Analysis of Surface States on Silicon," *Phys. Status Solidi*, **3**, 447 (1963).
9. E. H. Nicollian and J. R. Brews, *MOS Physics and Technology*, Wiley, New York, 1982.

10. C. G. B. Garrett and W. H. Brattain, "Physical Theory of Semiconductor Surfaces," *Phys. Rev.*, **99**, 376 (1955).
11. S. R. Hofstein and G. Warfield, "Physical Limitation on the Frequency Response of a Semiconductor Surface Inversion Layer," *Solid-State Electron.*, **8**, 321 (1965).
12. A. S. Grove, B. E. Deal, E. H. Snow, and C. T. Sah, "Investigation of Thermally Oxidized Silicon Surfaces Using Metal-Oxide-Semiconductor Structures," *Solid-State Electron.*, **8**, 145 (1965).
13. A. S. Grove, E. H. Snow, B. E. Deal, and C. T. Sah, "Simple Physical Model for the Space-Charge Capacitance of Metal-Oxide-Semiconductor Structures," *J. Appl. Phys.*, **33**, 2458 (1964).
14. J. R. Brews, "A Simplified High-Frequency MOS Capacitance Formula," *Solid-State Electron.*, **20**, 607 (1977).
15. A. Goetzberger, "Ideal MOS Curves for Silicon," *Bell Syst. Tech. J.*, **45**, 1097 (1966).
16. B. E. Deal, "Standardized Terminology for Oxide Charges Associated with Thermally Oxidized Silicon," *IEEE Trans. Electron Dev.*, **ED-27**, 606 (1980).
17. I. Tamm, "Über eine mögliche Art der Elektronenbindung an Kristalloberflächen," *Phys. Z. Sowjetunion,* **1**, 733 (1933).
18. W. Shockley, "On the Surface States Associated with a Periodic Potential," *Phys. Rev.*, **56**, 317 (1939).
19. W. Shockley and G. L. Pearson, "Modulation of Conductance of Thin Films of Semiconductors by Surface Charges," *Phys. Rev.*, **74**, 232 (1948).
20. F. G. Allen and G. W. Gobeli, "Work Function, Photoelectric Threshold and Surface States of Atomically Clean Silicon," *Phys. Rev.*, **127**, 150 (1962).
21. E. H. Nicollian and A. Goetzberger, "MOS Conductance Technique for Measuring Surface State Parameters," *Appl. Phys. Lett.*, **7**, 216 (1965).
22. C. N. Berglund, "Surface States at Steam-Grown Silicon-Silicon Dioxide Interface," *IEEE Trans. Electron Dev.*, **ED-13**, 701 (1966).
23. R. Castagne and A. Vapaille, "Description of the SiO_2-Si Interface Properties by Means of Very Low Frequency MOS Capacitance Measurements," *Surface Sci.*, **28**, 157 (1971).
24. E. H. Nicollian and A. Goetzberger, "The Si-SiO_2 Interface-Electrical Properties as Determined by the MIS Conductance Technique," *Bell Syst. Tech. J.*, **46**, 1055 (1967).
25. M. H. White and J. R. Cricchi, "Characterization of Thin-Oxide MNOS Memory Transistors," *IEEE Trans. Electron Dev.*, **ED-19**, 1280 (1972).
26. B. E. Deal, M. Sklar, A. S. Grove, and E. H. Snow, "Characteristics of the Surface-State Charge (Q_{ss}) of Thermally Oxidized Silicon," *J. Electrochem. Soc.*, **114**, 266 (1967).
27. J. R. Ligenza, "Effect of Crystal Orientation on Oxidation Rates of Silicon in High Pressure Steam," *J. Phys. Chem.*, **65**, 2011 (1961).
28. E. H. Snow, A. S. Grove, B. E. Deal, and C. T. Sah, "Ion Transport Phenomena in Insulating Films," *J. Appl. Phys.*, **36**, 1664 (1965).
29. B. E. Deal, E. H. Snow, and C. A. Mead, "Barrier Energies in Metal-Silicon Dioxide-Silicon Structures," *J. Phys. Chem. Solids*, **27**, 1873 (1966).
30. R. Williams, "Photoemission of Electrons from Silicon into Silicon Dioxide," *Phys. Rev.*, **140**, A569 (1965).
31. K. L. Jensen, "Electron Emission Theory and its Application: Fowler-Nordheim Equation and Beyond," *J. Vac. Sci. Technol. B,* **21**, 1528 (2003).

32. J. Frenkel, "On the Theory of Electric Breakdown of Dielectrics and Electronic Semiconductors," *Tech. Phys. USSR*, **5**, 685 (1938); "On Pre-Breakdown Phenomena in Insulators and Electronic Semiconductors," *Phys. Rev.*, **54**, 647 (1938).
33. Y. Takahashi and K. Ohnishi, "Estimation of Insulation Layer Conductance in MNOS Structure," *IEEE Trans. Electron Dev.*, **ED-40**, 2006 (1993).
34. J. J. O'Dwyer, *The Theory of Electrical Conduction and Breakdown in Solid Dielectrics*, Clarendon, Oxford, 1973.
35. S. M. Sze, "Current Transport and Maximum Dielectric Strength of Silicon Nitride Films:" *J. Appl. Phys.*, **38**, 2951 (1967).
36. W. C. Johnson, "Study of Electronic Transport and Breakdown in Thin Insulating Films," *Tech. Rep.* No.7, Princeton University, 1979.
37. M. Av-Ron, M. Shatzkes, T. H. DiStefano, and I. B. Cadoff, "The Nature of Electron Tunneling in SiO_2," in S. T. Pantelider, Ed., *The Physics of SiO_2 and Its Interfaces*, Pergamon, New York, 1978, p. 46.
38. A. Rusu and C. Bulucea, "Deep-Depletion Breakdown Voltage of SiO_2/Si MOS Capacitors," *IEEE Trans. Electron Dev.*, **ED-26**, 201 (1979).
39. E. H. Nicollian, A. Goetzberger, and C. N. Berglund, "Avalanche Injection Currents and Charging Phenomena in Thermal SiO_2," *Appl. Phys. Lett.*, **15**, 174 (1969).
40. D. R. Collins and C. T. Sah, "Effects of X-Ray Irradiation on the Characteristics of MOS Structures," *Appl. Phys. Lett.*, **8**, 124 (1966).
41. E. H. Snow, A. S. Grove, and D. J. Fitzgerald, "Effect of Ionization Radiation on Oxidized Silicon Surfaces and Planar Devices," *Proc. IEEE*, **55**, 1168 (1967).
42. J. Colinge and C. A. Colinge, *Physics of Semiconductor Devices*, Kluwer, Boston, 2002.
43. Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H. C. Wann, S. J. Wind, and H. Wong, "CMOS Scaling into the Nanometer Regime" *Proc. IEEE*, **85**, 486 (1997).
44. J. S. Suehle, "Ultrathin Gate Oxide Reliability: Physical Models, Statistics, and Characterization," *IEEE Trans. Electron Dev.*, **ED-49**, 958 (2002).
45. J. H. Stathis, "Physical and Predictive Models of Ultrathin Oxide Reliability in CMOS Devices and Circuits," *IEEE Trans. Device Mater. Reliab.*, **1**, 43 (2001).
46. J. S. Suehle and P. Chaparala, "Low Electric Field Breakdown of Thin SiO_2 Films Under Static and Dynamic Stress," *IEEE Trans. Electron Dev.*, **ED-44**, 801 (1997).

PROBLEMS

1. For an ideal Si-SiO_2 MOS capacitor with $d = 10$ nm, $N_A = 5\times10^{17}$ cm^{-3}, find the applied voltage and the electric fields at the SiO_2-Si interface required (a) to make the silicon surface intrinsic, and (b) to bring about a strong inversion.
2. Plot the variation of the space charge density $|Q_s|$ as a function of the surface potential ψ_s for an n-type silicon with $N_D = 10^{16}$ cm^{-3} at 300 K. Refer to Fig. 5 (p. 203). On the plot, mark the value of $2\psi_B$, and the magnitude of Q_s at the onset of strong inversion.
3. Derive the differential capacitance of the semiconductor depletion layer at the flat-band condition. (Eq. 23).

4. Derive an expression for the approximated segment of an ideal MOS C-V curve in the depletion case (i.e., $0 \le V \le V_T$ in Fig. 7, p. 206).
(Hint: The expression is either

$$\frac{C}{C_i} = \frac{1}{\sqrt{1 + \gamma V}} \quad \text{or} \quad \frac{C}{C_i} = \frac{1}{1 + \sqrt{\gamma V}}$$

where $\gamma \equiv 2\varepsilon_i^2 / qN_A\varepsilon_s d^2$ and V is the applied voltage on the metal plate.)

5. Find the charge per unit area in the inversion region for an ideal MOS capacitor with $N_A = 10^{16}$ cm^{-3}, $d = 10$ nm, and $V_G = 1.77$ V.

6. For a metal-SiO$_2$-Si capacitor having $N_A = 10^{16}$ cm^{-3} and $d = 8$ nm, calculate the minimum capacitance on the C-V curve under high-frequency condition.

7. An ideal Si MOS capacitor has an oxide of 5 nm and a doping of $N_A = 10^{17}$ cm^{-3}. Find the width of the inversion region, when the surface potential is 10% larger than the potential difference between the Fermi level and the intrinsic Fermi level.

8. Plot the number of electrons per unit area in the inversion region (N_I) of a silicon MOS capacitor versus surface electric field ($\mathscr{E}_s$). The substrate doping is 10^{17} cm^{-3}. Use log-log plot covering N_I from 10^9 to 10^{13} cm^{-2} and $\mathscr{E}_s$ from 10^5 to 10^6 V/cm. Also write down the value of N_I for $\mathscr{E}_s = 2.5\times10^5$ V/cm.

9. An ideal silicon MOS (MO-p-π-p^+) capacitor has an oxide thickness of 100 nm and a special doping profile of p-π-p^+ where the top p-layer is 10^{16} cm^{-3} and 1.5 μm thick and the π-layer is 3 μm thick. Find the breakdown voltage of the structure under pulse condition.

10. Plot an ideal C-V curve for a Si-SiO$_2$ MOS capacitor at 300 K with $N_A = 5\times10^{15}$ cm^{-3}, $d =$ 3 nm (specify C_i, $C_{\min}$, C_{FB}, and V_T). If the metal work function is 4.5 eV, $q\chi = 4.05$ eV, $Q_f/q = 10^{11}$ cm ^{2}, $Q_m/q = 10^{10}$ cm^{-2}, $Q_{ot}/q = 5\times10^{10}$ cm^{-2}, and $Q_{it} = 0$, plot the corresponding C-V curve (specify V_{FB} and the new V_T).

11. From the high-field portion in Fig. 25 (p. 230), evaluate the dielectric constant of the material.

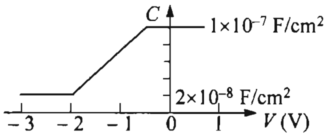
12. Assume that the oxide trapped charge Q_{ot} in an oxide layer is a charge sheet with an area density of 5×10^{11} cm^{-2} located at $y = 5$ nm from the metal-oxide interface. The thickness of the oxide layer is 10 nm. Find the change in the flat-band voltage due to Q_{ot}.

13. Derive Eqs. 39 and 40. Find the maximum value of G/ω.

14. Two MOS capacitors, both have 15 nm gate oxide. One has an n^+-polysilicon gate and p-type substrate, another has a p^+-polysilicon gate and n-type substrate. If the threshold voltage of these two capacitors are $V_{Tn} = |V_{Tp}| = 0.5$ V, and $Q_f = Q_m = Q_{ot} = Q_{it} = 0$, find the substrate dopings N_A and N_D.

15. (a) Calculate the change in flat-band voltage corresponding to a uniform positive charge distribution in the oxide. The total density of ions is 10^{12} cm^{-2} and the oxide thickness is 0.2 μm.
    (b) Calculate the change in V_{FB} for the same total density of ions and same oxide thickness as in (a) except that the charge has a triangular distribution which is high near the metal and zero near the silicon.

16. The C-V curve of a Si MOS capacitor is shown in the right figure. The shift is due entirely to the fixed oxide charges at the SiO_2-Si interface. It has an n^+-poly gate. Find the number of fixed oxide charges.



17. Based on the plot for weak inversion region on p. 222 (Fig. 18), find the resistance associated with the interface traps.

18. An MOS capacitor has an oxide of 10 nm and a substrate doping of $N_A = 10^{16}$ cm^{-3}. The capacitor has a positive gate bias of 2 V and a surface potential of 0.91 V. When the capacitor is illuminated, an additional charge sheet of 10^{12} electrons/cm^2 is formed at the SiO_2-Si interface. Calculate the percentage change of the high-frequency capacitance, i.e.,

$$\frac{C(\text{under illumination})}{C(\text{no illumination})} - 1 .$$

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

6

MOSFETs

6.1 INTRODUCTION
6.2 BASIC DEVICE CHARACTERISTICS
6.3 NONUNIFORM DOPING AND BURIED-CHANNEL DEVICE
6.4 DEVICE SCALING AND SHORT-CHANNEL EFFECTS
6.5 MOSFET STRUCTURES
6.6 CIRCUIT APPLICATIONS
6.7 NONVOLATILE MEMORY DEVICES
6.8 SINGLE-ELECTRON TRANSISTOR

6.1 INTRODUCTION

The metal-oxide-semiconductor field-effect transistor (MOSFET) is the most-important device for forefront high-density integrated circuits such as microprocessors and semiconductor memories. It is also becoming an important power device. The principle of the surface field-effect transistor was first proposed in the early 1930s by Lilienfeld[1–3] and Heil.[4] It was subsequently studied by Shockley and Pearson[5] in the late 1940s. In 1960, Ligenza and Spitzer produced the first device-quality Si-SiO_2 MOS system using thermal oxidation.[6] The basic MOSFET structure using this Si-SiO_2 system was proposed by Atalla.[7] Subsequently the first MOSFET was reported by Kahng and Atalla in 1960.[8] The detailed early historical development of the MOSFET can be found in Refs. 9–10. The basic device characteristics have been initially studied by Ihantola and Moll,[11] Sah,[12] and Hofstein and Heiman.[13] The technology, application, and device physics have been reviewed by many books.[14–17]

Figure 1 shows the reduction of the gate-length dimension in production ICs since 1970. This dimension has been decreasing at a steady pace and will continue to shrink in the foreseeable future. The reduction of device dimensions is driven by the require-

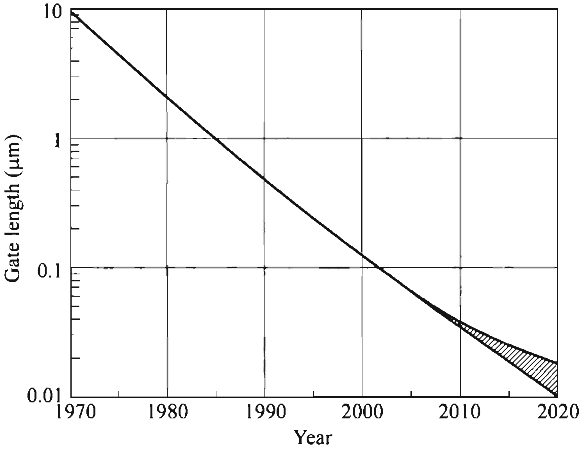


Fig. 1 Minimum gate dimension in commercial integrated circuit as a function of the year of production.

ment for both performance and density. The number of components per integrated-circuit chip has grown exponentially. The rate of growth is expected to slow down because of increasing technological challenge and fabrication cost. However, a complexity of 1 billion or more devices per chip had been available around 2000 using 0.1-μm technology.

In this chapter we first consider the basic device characteristics of the so-called long-channel MOSFET; that is, the longitudinal field along the channel is not large enough to cause velocity saturation. In this regime, the carrier velocity is mobility-limited, or under constant mobility. As the channel length becomes shorter, one has to consider short-channel effects due to two-dimensional potential and high-field transport such as velocity saturation and ballistic transport. Many device structures have been proposed to improve MOSFET performance. Some representative advanced structures as well as the nonvolatile semiconductor memory, basically a MOSFET with a multilayer gate structure, will be discussed.

6.1.1 Field-Effect Transistors: Family Tree

The MOSFET is the main member of the family of field-effect transistors. A distinction between the field-effect transistor (FET) and the potential-effect transistor (PET) is warranted here. A transistor in general is a three-terminal device where the channel resistance between two of the contacts is controlled by the third (MOSFETs have the fourth terminal as contact to the substrate). The difference between the FET and the PET is the way the control is coupled to the channel. As shown in Fig. 2, in an FET, the channel is controlled *capacitively* by an electric field (hence the name *field-effect*), and in a PET, the channel's potential is accessed directly (hence the name *potential-effect*). Conventionally in FETs, the channel carriers flow from the source

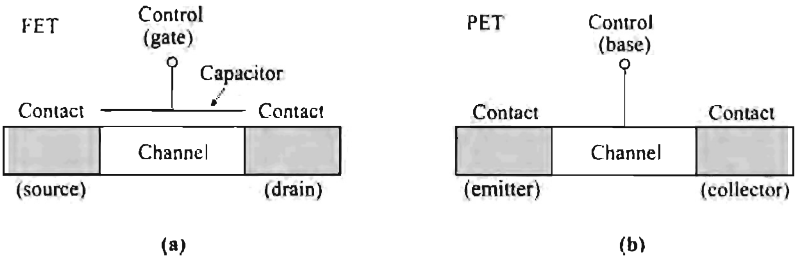

FET
Control
(gate)
Capacitor
Contact
Contact
Channel
(source)
(drain)
(a)
PET
Control
(base)
Contact
Contact
Channel
(emitter)
(collector)
(b)

Fig. 2 Distinction between (a) field-effect transistor (FET) and (b) potential-effect transistor (PET).

to the drain, and the control terminal is called the gate, whereas in PETs, these corresponding terminals are called the emitter, collector, and base, respectively. The bipolar transistor is a good representative of the PETs.

A family tree of field-effect transistors is shown in Fig. 3 The three first-level main members are IGFET (insulated-gate FET), JFET (junction FET), and MESFET (metal-semiconductor FET). They are distinguished by the way the gate capacitor is formed. In an IGFET, the gate capacitor is an insulator. In a JFET or a MESFET, the capacitor is formed by the depletion layer of a *p-n* junction or a Schottky barrier, respectively. In the branch of IGFET, we further divide it into MOSFET/MISFET (metal-insulator-semiconductor FET) and HFET (heterojunction FET). In the MOSFET, specifically the insulator is a grown oxide layer, whereas in the MISFET the insulator is a deposited dielectric. In the HFET branch, the gate material is a high-bandgap semiconductor layer grown as a heterojunction which acts as an insulator. Although MOSFETs have been made with various semiconductors such as Ge,[18] Si, and GaAs,[19] and use various oxides and insulators such as SiO_2, Si_3N_4, and Al_2O_3, the most-important system is the SiO_2-Si combination. Hence most of the results in this chapter are obtained from the SiO_2-Si system. The other members, JFETs, MESFETs, and HFETs, will be considered in the following chapter.

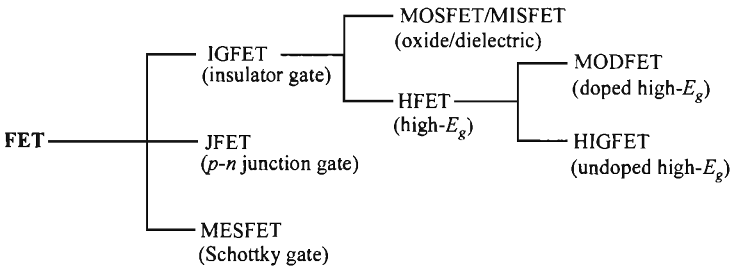

FET
IGFET
(insulator gate)
JFET
(p-n junction gate)
MESFET
(Schottky gate)
MOSFET/MISFET
(oxide/dielectric)
HFET
(high-E_g)
MODFET
(doped high-E_g)
HIGFET
(undoped high-E_g)

Fig. 3 Family tree of field-effect transistors (FETs).

Field-effect transistors offer many attractive features for applications in analog switching, high-input-impedance amplifiers, and microwave amplifiers, in addition to digital integrated circuits. The FETs have considerably higher input impedance than bipolar transistors, which allows the input of a FET to be more readily matched to the standard microwave system. The FET has a negative temperature coefficient at high current levels; that is, the current decreases as temperature increases. This characteristic leads to a more uniform temperature distribution over the device area and prevents the FET from thermal runaway or second breakdown, that can occur in the bipolar transistor. The device is thermally stable, even when the active area is large or when many devices are connected in parallel. Because there is no forward-biased *p-n* junctions, FETs do not suffer from minority-carrier storage and, consequently, have higher large-signal switching speeds. In addition, the devices are basically square-law or linear devices; intermodulation and cross-modulation products are smaller than those of bipolar transistors.

6.1.2 Versions of Field-Effect Transistors

There are many ways to categorize the versions of FETs. First, according to the type of channel carriers, we have *n*-channel and *p*-channel devices. *n*-channels are formed by electrons and are more conductive with more positive gate bias, while *p*-channels are formed by holes and are more conductive with more negative gate bias. Furthermore, it is important to describe the state of the transistor with zero gate bias. FETs are called enhancement-mode, or normally-off, if at zero gate bias the channel conductance is very low and we must apply a gate voltage to form a conductive channel. The counterpart is called depletion-mode, or normally-on, when the channel is conductive with zero gate bias and we must apply a gate voltage to turn the transistor off. These four combinations with their *I-V* characteristics are summarized in Fig. 4.

It is important also to point out the nature of the channel in more details. According to Fig. 5, a channel can be formed by a surface inversion layer or a bulk buried layer. The surface inversion channel is a two-dimensional charge sheet of thickness in the order of 5 nm. The buried channel is much thicker, comparable to the depletion width since when the transistor is turned off, the channel is totally consumed by the surface depletion layer. In the FET family, MESFETs and JFETs are always buried-channel devices, while MODFETs are surface-channel devices. MOSFETs and MISFETs can have both kinds of channels in parallel, but in practice, they are mostly surface-channel devices.

These two kinds of channels offer advantages of their own. Buried channels are based on bulk conduction and, thus, are free of surface effects such as scattering and surface defects, resulting in better carrier mobility. On the other hand, the physical distance between the gate and the channel is larger and also dependent on gate bias, leading to a lower and variable transconductance. Note that for depletion-mode devices, it is common to use buried channels, but theoretically, one can achieve the same goal by choosing a gate material with a proper work function to shift the threshold voltage to a desirable value.

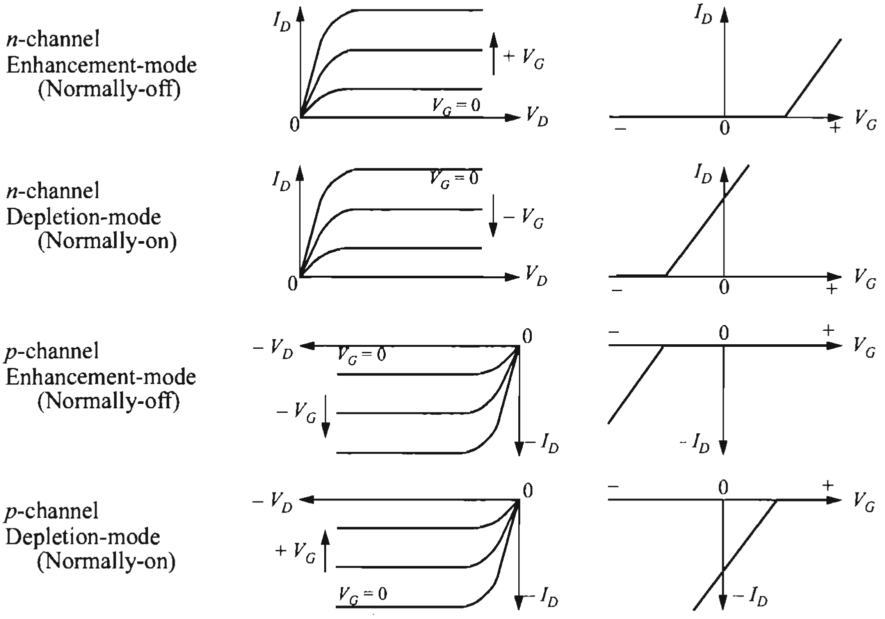


Fig. 4 Versions of MOSFETs; their output and transfer characteristics.

6.2 BASIC DEVICE CHARACTERISTICS

The basic structure of a MOSFET is illustrated in Fig. 6. Throughout this chapter we assume the channel carriers are electrons—an *n*-channel device. All discussion and equations will be applicable to the counterpart *p*-channel devices with appropriate substitution of parameters and the reversal of polarity of the applied voltages. A common MOSFET is a four-terminal device that consists of a *p*-type semiconductor substrate into which two n^+-regions, the source and drain, are formed, usually by ion implantation. The SiO_2 gate dielectric is formed by thermal oxidation of Si for a high-quality SiO_2-Si interface. The metal contact on the insulator is called the gate; heavily

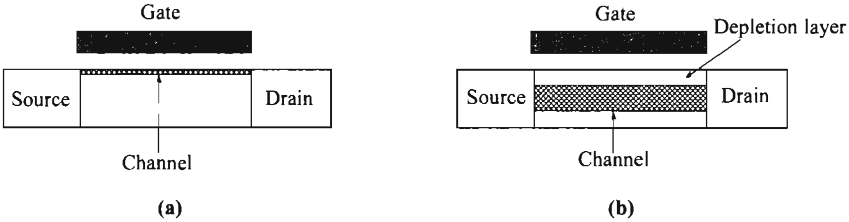


Fig. 5 FET channels: (a) surface inversion channel and (b) buried channel.

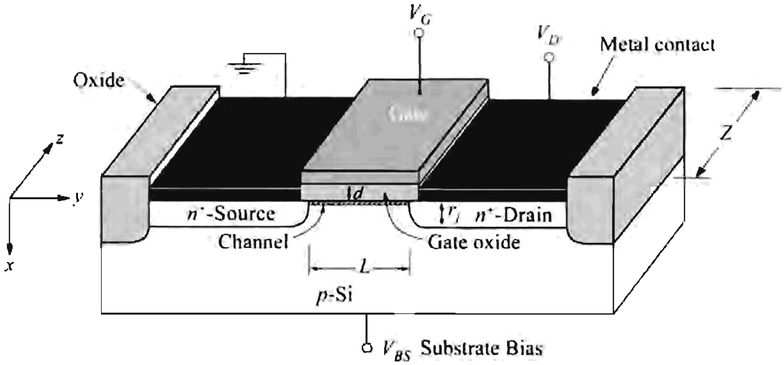


Fig. 6 Schematic diagram of a MOSFET.

doped polysilicon or a combination of silicide and polysilicon is more commonly used as the gate electrode. The basic device parameters are the channel length L, which is the distance between the two metallurgical n^+-p junctions; the channel width Z; the insulator thickness d; the junction depth r_j; and the substrate doping N_A. In a silicon integrated circuit, a MOSFET is surrounded by a thick oxide (called the field oxide to distinguish it from the gate oxide) or a trench filled with insulator to electrically isolate it from adjacent devices.

The source contact will be used as the voltage reference throughout this chapter. When ground or a low voltage is applied to the gate, the main channel is shut off, and the source-to-drain electrodes correspond to two p-n junctions connected back to back. When a sufficiently large positive bias is applied to the gate so that a surface inversion layer (or channel) is formed between the two n^+-regions, the source and the drain are then connected by a conducting surface n-channel through which a large current can flow. The conductance of this channel can be modulated by varying the gate voltage. The back-surface contact (or substrate contact) can be at the reference voltage or reverse biased; this substrate voltage will also affect the channel conductance.

6.2.1 Inversion Charge in Channel

When a voltage is applied across the source-drain contacts, the MOS structure is in a nonequilibrium condition; that is, the minority-carrier (electron in the present case) quasi-Fermi level E_{Fn} is lowered from the equilibrium Fermi level. To show more clearly the band bending across the device, Fig. 7a shows the MOSFET turned 90°. The two-dimensional, flat-band, zero-bias ($V_G = V_D = V_{BS} = 0$) equilibrium condition is shown in Fig. 7b. The equilibrium condition but under a gate bias that causes surface inversion is shown in Fig. 7c. The nonequilibrium condition with both drain and gate biases is shown in Fig. 7d, where we note the separation of the quasi-Fermi levels of electrons E_{Fn} and holes E_{Fp}; the E_{Fp} remains at the bulk Fermi level while

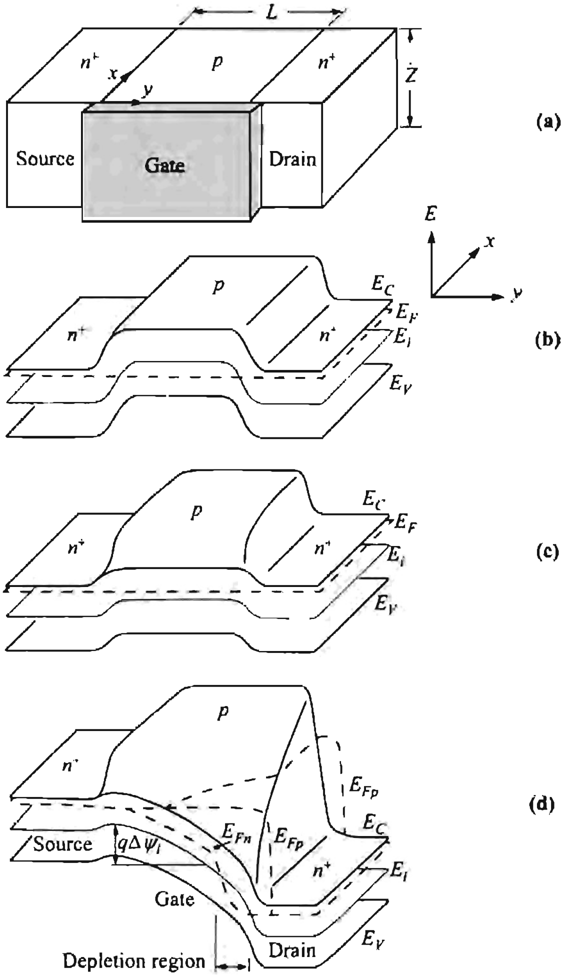

L
n+
p
n+
x
y
Z
Source
Gate
Drain
(a)
E
x
y
p
n+
n+
E_C
E_F
E_i
E_V
(b)
p
n+
n+
E_C
E_F
E_i
E_V
(c)
p
n+
Source
qΔψ_i
E_Fn
E_Fp
E_Fp
E_C
n+
E_i
Gate
Depletion region
Drain
E_V
(d)

Fig. 7 Two-dimensional band diagram of an *n*-channel MOSFET. (a) Device configuration. (b) Flat-band zero-bias equilibrium condition. (c) Equilibrium condition ($V_D = 0$) under a positive gate bias. (d) Nonequilibrium condition under both gate and drain biases. (After Ref. 20.)

E_{Fn} is lowered toward the drain contact. Figure 7d shows that the gate voltage required for inversion at the drain is larger than the equilibrium case in which $\psi_s(\text{inv}) \approx 2\psi_B$;# in other words, the inversion-layer charge at the drain end is lowered by the drain bias. This is because the applied drain bias lowers the E_{Fn}, and an inversion

---

The common assumption of $\psi_s = 2\psi_B$ is for the onset of weak inversion. For strong inversion, ψ_s can be larger by a few kT.[15] This can be understood from Fig. 5 of Chapter 4.

layer can be formed only when the surface potential meets the criteria of $[E_{Fn} - E_i(0)] > q\psi_B$, where $E_i(0)$ is the intrinsic Fermi level at $x = 0$.

Figure 8 shows a comparison of the charge distribution and energy-band variation of an inverted p-region for the equilibrium case and the nonequilibrium case at the drain. For the equilibrium case, the surface depletion region reaches a maximum width W_{Dm} at inversion. For the nonequilibrium case, the depletion-layer width is deeper than W_{Dm} and is a function of the drain bias V_D. The surface potential $\psi_s(y)$ at the drain at the onset of strong inversion is, to a good approximation, given by

$$\psi_s(\text{inv}) \approx V_D + 2\psi_B . \tag{1}$$

The characteristics of the surface space charge under the nonequilibrium condition are derived under two assumptions; (1) the majority-carrier quasi-Fermi level E_{Fp} is the same as that of the substrate and it does not vary with distance from the bulk to the surface (constant with x), and (2) the minority-carrier quasi-Fermi level E_{Fn} is lowered by the drain bias by an amount dependent on the y-position. The first assumption introduces little error when the surface is inverted, because majority carriers are then only a negligible part of the surface space charge. The second assumption is correct under the inversion condition, because minority carriers are an important part of the surface space-charge region when the surface is inverted.

Based on these assumptions, the one-dimensional Poisson equation for the surface space-charge region at the drain end is given by

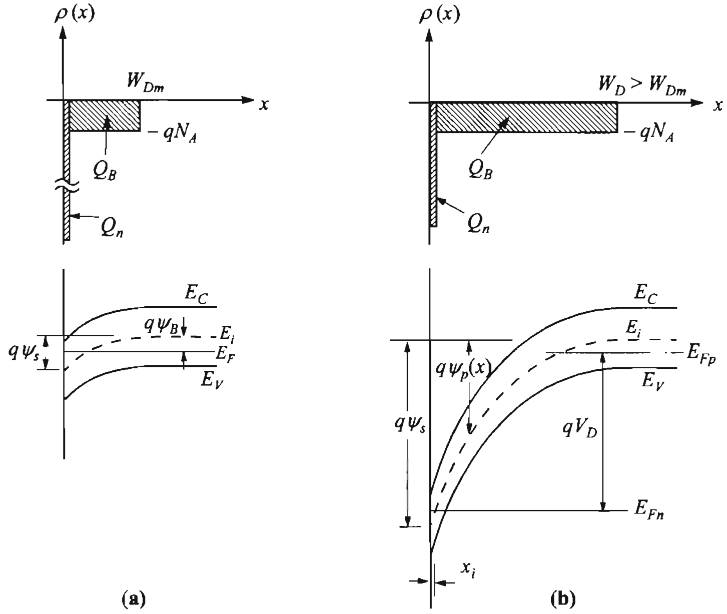


Fig. 8 Comparison of charge distribution and energy-band variation of an inverted p-region in (a) equilibrium and (b) nonequilibrium at the drain end. (After Ref. 21.)

$$\frac{d^2\psi_p}{dx^2} = \frac{q}{\varepsilon_s}(N_A - p + n) \tag{2}$$

where

$$p_{po} = N_A = \frac{n_i^2}{n_{po}} \tag{3}$$

$$p = N_A \exp(-\beta\psi_p) \tag{4}$$

$$n = n_{po}\exp(\beta\psi_p - \beta V_D), \tag{5}$$

and $\beta \equiv q/kT$.

Conceptually the charge due to minority carriers within the inversion layer, is given by

$$\begin{aligned}|Q_n| &\equiv q\int_0^{x_i} n(x)dx = q\int_{\psi_s}^{\psi_B} \frac{n(\psi_p)d\psi_p}{d\psi_p/dx} \\ &= q\int_{\psi_s}^{\psi_B} \frac{n_{po}\exp(\beta\psi_p - \beta V_D)d\psi_p}{(\sqrt{2}kT/qL_D)F(\beta\psi_p, V_D, n_{po}/p_{po})}\end{aligned} \tag{6}$$

where x_i denotes the point at which $q\psi_p(x) = E_{Fn} - E_i(x) = q\psi_B$, and the function F is defined as (see Chapter 4)

$$F\left(\beta\psi_p, V_D, \frac{n_{po}}{p_{po}}\right) \equiv$$

$$\sqrt{\exp(-\beta\psi_p) + \beta\psi_p - 1 + \frac{n_{po}}{p_{po}}\exp(-\beta V_D)[\exp(\beta\psi_p) - \beta\psi_p\exp(\beta V_D) - 1]}. \tag{7}$$

For the practical doping ranges in silicon, the value of x_i is quite small, of the order of 3 to 30 nm. Equation 6 is the exact formulation, but can only be evaluated numerically.

To get an analytical solution we follow the same approach as in Chapter 4 (Eq. 14). The surface electric field in the x-direction at the drain end is given by

$$\mathscr{E}_s = -\left.\frac{d\psi_p}{dx}\right|_{x=0} = \pm\frac{\sqrt{2}kT}{qL_D}F\left(\beta\psi_s, V_D, \frac{n_{po}}{p_{po}}\right), \tag{8}$$

and the total semiconductor surface charge is then obtained from Gauss' law

$$Q_s = -\varepsilon_s\mathscr{E}_s = \mp\frac{\sqrt{2}\varepsilon_s kT}{qL_D}F\left(\beta\psi_s, V_D, \frac{n_{po}}{p_{po}}\right), \tag{9}$$

where the Debye length is

$$L_D \equiv \sqrt{\frac{kT\varepsilon_s}{N_A q^2}}. \tag{10}$$

The inversion charge per unit area Q_n after strong inversion is given by

$$Q_n = Q_s - Q_B \tag{11}$$

where the depletion bulk charge is

$$Q_B = -qN_AW_D = -\sqrt{2qN_A\varepsilon_s(V_D + 2\psi_B)}. \tag{12}$$

From Eqs. 9, 11, and 12, the inversion charge Q_n at the drain end can be simplified to

$$|Q_n| \approx \sqrt{2}qN_AL_D\left[\sqrt{\beta\psi_s + \left(\frac{n_{po}}{p_{po}}\right)\exp(\beta\psi_s - \beta V_D)} - \sqrt{\beta\psi_s}\right]. \tag{13}$$

This solution is still difficult to use because at strong inversion, Q_n is very sensitive to the surface potential ψ_s (see Fig. 5 in Chapter 4). Another shortcoming is that the relationship to the terminal bias, that is V_G, is still missing. The charge-sheet model discussed in the following section is simpler and much more useful for deriving the I-V characteristics of MOSFETs.

Charge-Sheet Model. In the charge-sheet model,[22] under strong-inversion conditions, the inversion layer is treated as a charge sheet with zero thickness ($x_i = 0$). Consequently, this assumption implies that the potential drop across this charge sheet is also zero. These assumptions do introduce error but within an acceptable level. From Gauss' law, the boundary conditions on both sides of the charge sheet are:

$$\mathscr{E}_{ox}\varepsilon_{ox} = \mathscr{E}_s\varepsilon_s - Q_n. \tag{14}$$

In order to express $Q_n(y)$ throughout the channel, the surface potential is generalized from Eq. 1 to

$$\psi_s(y) \approx \Delta\psi_i(y) + 2\psi_B, \tag{15}$$

where $\Delta\psi_i$ is the channel potential with respect to the source end;

$$\Delta\psi_i(y) \equiv \frac{E_i(x=0,y=0) - E_i(x=0,y)}{q}, \tag{16}$$

(see label in Fig. 7d) and is equal to V_D at the drain end. Note that the electric fields can be expressed as:

$$\mathscr{E}_{ox} = \frac{V_G - \psi_s}{d} = \frac{V_G - (\Delta\psi_i + 2\psi_B)}{d}, \tag{17}$$

$$\mathscr{E}_s = \sqrt{\frac{2qN_A(\Delta\psi_i + 2\psi_B)}{\varepsilon_s}}. \tag{18}$$

In Eq. 17, an ideal MOS system with zero work-function difference is assumed. Equation 18 is simply the maximum field at the edge of the depletion region. Combining Eqs. 14–18 and using $C_{ox} = \varepsilon_{ox}/d$ we obtain

$$|Q_n(y)| = [V_G - \Delta\psi_i(y) - 2\psi_B]C_{ox} - \sqrt{2\varepsilon_sqN_A[\Delta\psi_i(y) + 2\psi_B]}. \tag{19}$$

This final form will be used as the channel charge responsible for the current conduction.

6.2.2 Current-Voltage Characteristics

We shall now derive the basic MOSFET characteristics under the following idealized conditions: (1) The gate structure corresponds to an ideal MOS capacitor as defined in Chapter 4; that is, there are no interface traps nor mobile oxide charge; (2) only drift current will be considered; (3) doping in the channel is uniform; (4) reverse leakage current is negligible; and (5) the transverse field ($\mathscr{E}_x$ in the x-direction) in the channel is much larger than the longitudinal field ($\mathscr{E}_y$ in the y-direction). This last condition corresponds to the so-called gradual-channel approximation. Note that in condition-(1), the requirements of zero fixed oxide charge and work-function difference are removed, and their effects are included in a flat-band voltage V_{FB} required by the gate to produce the flat-band condition. Consequently V_G is replaced by $V_G - V_{FB}$ for the inversion charge, giving

$$|Q_n(y)| = [V_G - V_{FB} - \Delta\psi_i(y) - 2\psi_B]C_{ox} - \sqrt{2\varepsilon_s q N_A[\Delta\psi_i(y) + 2\psi_B]}\,. \tag{20}$$

Under such idealized conditions, the channel current at any y-position is given by

$$I_D(y) = Z|Q_n(y)|v(y) \tag{21}$$

where $v(y)$ is the average carrier velocity. Since the current has to be continuous and constant throughout the channel, integration of Eq. 21 from 0 to L gives

$$I_D = \frac{Z}{L}\int_0^L |Q_n(y)|v(y)dy\,. \tag{22}$$

The carrier velocity $v(y)$ is a function of the y-position since the longitudinal field $\mathscr{E}_y(y)$ is a variable. Because of this, the relationship between $v(y)$ and $\mathscr{E}_y(y)$ is important to evaluate Eq. 22. We first consider the case where $\mathscr{E}_y(y)$ is low such that the mobility is constant. For shorter channel lengths, higher field causes velocity saturation and ultimately ballistic transport. These interesting effects will be discussed later.

Constant Mobility. Under this assumption, substitutions of $v = \mathscr{E}\mu$ and Eq. 20 into Eq. 22 gives

$$I_D = \frac{Z\mu_n}{L}\int_0^L |Q_n(y)|\mathscr{E}(y)dy = \frac{Z\mu_n}{L}\int_0^L |Q_n(y)|\frac{d\Delta\psi_i(y)}{dy}dy = \frac{Z\mu_n}{L}\int_0^{V_D} |Q_n(\Delta\psi_i)|d\Delta\psi_i$$
$$= \frac{Z}{L}\mu_n C_{ox}\left\{\left(V_G - V_{FB} - 2\psi_B - \frac{V_D}{2}\right)V_D - \frac{2}{3}\frac{\sqrt{2\varepsilon_s q N_A}}{C_{ox}}[(V_D + 2\psi_B)^{3/2} - (2\psi_B)^{3/2}]\right\}. \tag{23}$$

Equation 23 predicts that for a given V_G the drain current first increases linearly with drain voltage (the linear region), then gradually levels off (the nonlinear region), and finally approaching a saturated value (the saturation region). The basic output characteristics of an idealized MOSFET are shown in Fig. 9. The dashed line on the right indicates the locus of the drain voltage ($V_{D\text{sat}}$) at which the current reaches a maximum value $I_{D\text{sat}}$. For small V_D, the I_D is linear with V_D. Inbetween the two dashed lines, we designate this as the nonlinear region.

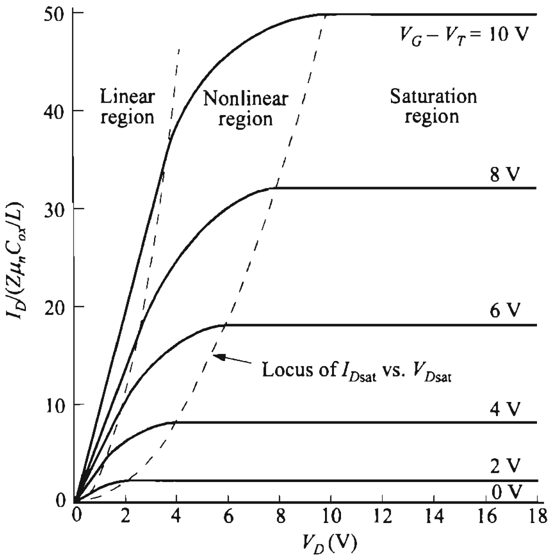


Fig. 9 Idealized drain characteristics (I_D vs. V_D) of a MOSFET. The dashed lines separate the linear, nonlinear, and saturation regions.

A qualitative discussion of the device operation can be helpful, with the aid of Fig. 10. Let us consider that a positive voltage is applied to the gate, large enough to cause an inversion at the semiconductor surface. If a small drain voltage is applied, a current will flow from the source to the drain through the conducting channel. The channel acts as a resistor, and the drain current I_D is proportional to the drain voltage V_D. This is the linear region. As the drain voltage increases, the current deviates from the linear relationship since the charge near the drain end is reduced by the channel potential $\Delta\psi_i$ (Eq. 20). It eventually reaches a point at which the inversion charge at the drain end $Q_n(L)$ is reduced to nearly zero. This location of $Q_n \approx 0$ is called the pinch-off point, Fig. 10b. [In reality $Q_n(L)$ is not zero for current continuity, but small because of its high field and high carrier velocity.] Beyond this drain bias, the drain current remains essentially the same, because for $V_D > V_{Dsat}$, the pinch-off point starts to move toward the source, but the voltage at this pinch-off point remains the same (V_{Dsat}). Thus, the number of carriers arriving at the pinch-off point from the source, and hence the current, remains essentially the same, apart from a decrease in L to the value L' (Fig. 10c). This change of effective channel length will increase the drain current only when the shortened amount is a substantial fraction of the channel length. This will be considered in the section of short-channel effects.

We shall now consider the current equations for the three cases of linear, nonlinear, and saturation regions. In the linear region, with a small V_D, using power series around V_D and taking only the initial terms, Eq. 23 reduces to

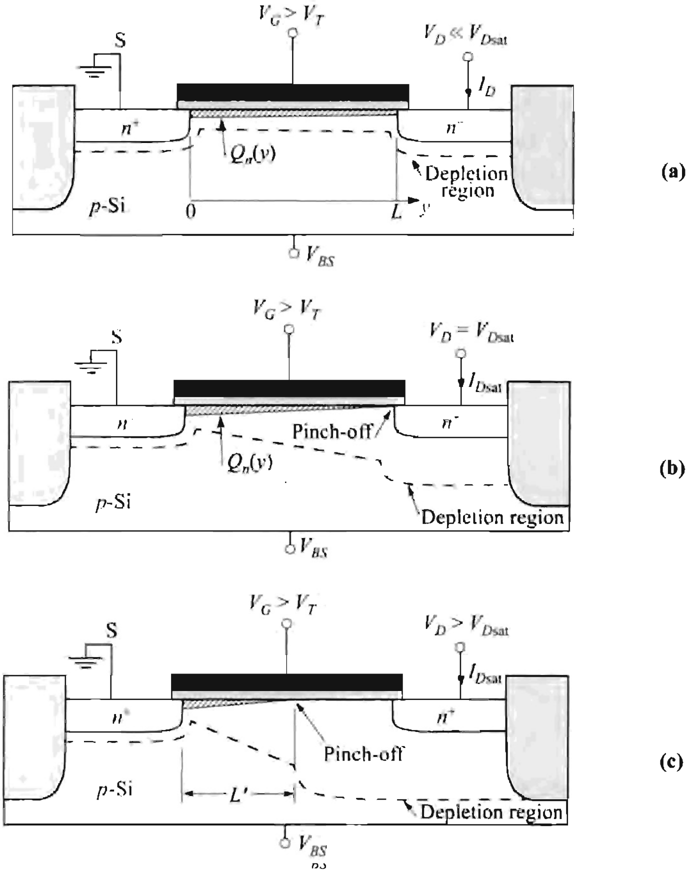

$V_G > V_T$
S
$V_D \ll V_{Dsat}$
I_D
n^+
n^+
$Q_n(y)$
Depletion region
p-Si
0
L
y
V_{BS}
(a)
$V_G > V_T$
S
$V_D = V_{Dsat}$
I_{Dsat}
n^+
n^+
Pinch-off
$Q_n(y)$
p-Si
Depletion region
V_{BS}
(b)
$V_G > V_T$
S
$V_D > V_{Dsat}$
I_{Dsat}
n^+
n^+
Pinch-off
p-Si
L'
Depletion region
V_{BS}
(c)

Fig. 10 MOSFET operated (a) in the linear region (low V_D), (b) at onset of saturation, and (c) beyond saturation (effective channel length is reduced).

$$
\begin{aligned}
I_D &= \frac{Z}{L}\mu_n C_{ox}\left\{\left(V_G - V_{FB} - 2\psi_B - \frac{V_D}{2}\right)V_D - \frac{2}{3}\frac{\sqrt{2\varepsilon_s q N_A}}{C_{ox}}\left(3\sqrt{\frac{\psi_B}{2}}V_D\right)\right\}\\
&= \frac{Z}{L}\mu_n C_{ox}\left(V_G - V_T - \frac{V_D}{2}\right)V_D \qquad \text{for} \quad V_D \ll (V_G - V_T)\ , \qquad (24)
\end{aligned}
$$

where V_T is the threshold voltage, one of the most-important parameters, given by

$$
V_T = V_{FB} + 2\psi_B + \frac{\sqrt{2\varepsilon_s q N_A(2\psi_B)}}{C_{ox}}\ . \qquad (25)
$$

The threshold voltage will be discussed in more details in the next section.

Careful examination of Eq. 23 indicates that the current initially increases but then goes through a peak and then drops with V_D. This drop of current is not physical, but it corresponds to the condition that the charge in the inversion layer at the drain end $Q_n(L)$ becomes zero. This pinch-off point occurs because the relative voltage between the gate and the semiconductor is reduced. The drain voltage and the drain current at this point are designated as $V_{D\text{sat}}$ and $I_{D\text{sat}}$, respectively. Beyond the pinch-off point the current remains independent of V_D and we have the saturation region. The value of $V_{D\text{sat}}$ is obtained from Eq. 20 under the condition $Q_n(L) = 0$. The solution yields

$$V_{D\text{sat}} = \Delta\psi_i(L) = V_G - V_{FB} - 2\psi_B + K^2\left[1 - \sqrt{1 + \frac{2(V_G - V_{FB})}{K^2}}\right] \tag{26}$$

where $K \equiv \sqrt{\varepsilon_s q N_A}/C_{ox}$. Alternatively, the same solution can be obtained by setting $dI_D/dV_D = 0$. The saturation current $I_{D\text{sat}}$ can be obtained by substituting Eq. 26 into Eq. 23:

$$I_{D\text{sat}} = \frac{Z}{2ML}\mu_n C_{ox}(V_G - V_T)^2. \tag{27}$$

M is a function of doping concentration and oxide thickness

$$M \equiv 1 + \frac{K}{2\sqrt{\psi_B}}. \tag{28}$$

It has a value slightly larger than unity and it approaches unity with thinner oxide and lower doping. Furthermore, a more convenient form for $V_{D\text{sat}}$ can be expressed as

$$V_{D\text{sat}} = \frac{V_G - V_T}{M}. \tag{29}$$

The transconductance in the saturation region where Eq. 27 applies is given by

$$g_m = \left.\frac{dI_D}{dV_G}\right|_{V_D > V_{D\text{sat}}} = \frac{Z}{ML}\mu_n C_{ox}(V_G - V_T). \tag{30}$$

It can be seen here that in this saturation region, for constant mobility, the current is a square-law function according to Eq. 27, indicated by the increasing current steps between gate bias shown in Fig. 9.

Finally, the nonlinear region inbetween these two extreme cases can be described well by

$$I_D = \frac{Z}{L}\mu_n C_{ox}\left(V_G - V_T - \frac{MV_D}{2}\right)V_D. \tag{31}$$

Equation 20 for the inversion charge is an exact expression. An approximation of the following form, taking advantage of the definition of threshold voltage, can be made:

$$|Q_n(y)| = C_{ox}[V_G - V_T - M\Delta\psi_i(y)]. \tag{32}$$

Substitution of this into Eq. 22 yields a general expression which is the same as Eq. 31 for the whole three regions of operation. As seen, the only slight departure from the previous results lies in the linear region. This simplified charge expression is helpful to analyze the conditions under field-dependent mobility and velocity saturation which is discussed next.

Velocity-Field Relationship. As technology advances and pushes for device performance and density, the channel length gets shorter and shorter. The internal longitudinal field $\mathscr{E}_y$ in the channel also increases as a result. The general υ-$\mathscr{E}$ relationship for high fields is shown in Fig. 11. Mobility μ is defined as $\upsilon/\mathscr{E}$. For low fields, the mobility is constant. This low-field mobility is used for the long-channel characteristics in the last section. In the extreme case of very high field, the velocity approaches a value, saturation velocity υ_s. Inbetween the constant-mobility regime and the saturation-velocity regime, the carrier velocity can be described by[23]

$$\upsilon(\mathscr{E}) = \frac{\mu_n \mathscr{E}}{[1+(\mu_n \mathscr{E}/\upsilon_s)^n]^{1/n}} = \frac{\mu_n \mathscr{E}}{[1+(\mathscr{E}/\mathscr{E}_c)^n]^{1/n}} \tag{33}$$

where μ_n is the low-field mobility. The value of n changes the shape of the curve, but μ_n, υ_s, and the critical field $\mathscr{E}_c$ $(\equiv \upsilon_s/\mu_n)$ remain the same. It has been observed that in silicon for electrons $n = 2$ and for holes $n = 1$ have the best fit. The value of υ_s for silicon at room temperature is around 1×10^7 cm/s.

As the terminal voltage V_D is increased from zero, current is increased because of higher field and higher velocity. Eventually the velocity reaches the maximum value of υ_s, and the current also saturates to a constant value. Notice that this current saturation comes from a completely different mechanism than in the case of constant mobility. Here, it is due to velocity saturation of carriers, before the pinch-off condition can occur.

To derive the I-V characteristics it is important to know the υ-$\mathscr{E}$ relationship (Fig. 11). We find that mathematically, for Eq. 33 with $n = 2$, the analysis is rather complicated. Fortunately for the cases of two-piece linear approximation and Eq. 33 with $n = 1$, the mathematics is manageable and simple solutions can be obtained. Since these two extremes mostly cover the realistic bounds for different kinds of carriers, we will consider both assumptions.

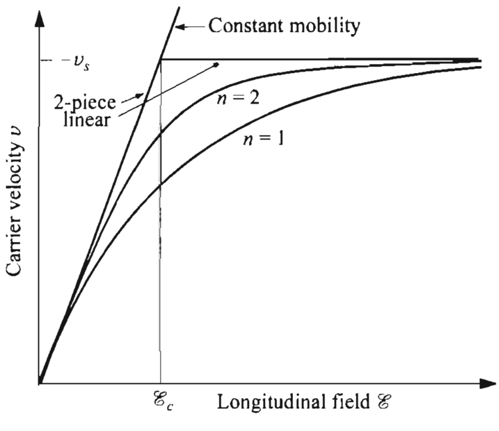


Fig. 11 υ-$\mathscr{E}$ relationship (Eq. 33) for $n = 1$ and 2, and two-piece linear approximation. The critical field $\mathscr{E}_c \equiv \upsilon_s/\mu$, where μ is low-field mobility, is also indicated.

Field-Dependent Mobility: Two-Piece Linear Approximation. In the two-piece linear approximation, the constant-mobility model is valid up to the point when the maximum field near the drain exceeds $\mathscr{E}_c$. Conversely, Eq. 23 is valid up to a new $V_{D\text{sat}}$ value which occurs earlier than the constant-mobility model, so the only task is to find that point. Substituting Eq. 32 into Eq. 21 gives

$$I_D(y) = ZC_{ox}\mu_n\mathscr{E}(V_G - V_T - M\Delta\psi_i). \tag{34}$$

Since we know the maximum field occurs at the drain, the current will saturate when the drain bias is increased to a value where $\mathscr{E}(L) = \mathscr{E}_c$. Equation 34 leads to the following condition:

$$I_{D\text{sat}} = ZC_{ox}\mu_n\mathscr{E}_c(V_G - V_T - MV_{D\text{sat}}). \tag{35}$$

Here we need one more equation to solve for two unknowns. Using Eqs. 32 and 22, we obtain an expression similar to Eq. 31 given as

$$I_{D\text{sat}} = \frac{ZC_{ox}\mu_n}{L}\left(V_G - V_T - \frac{MV_{D\text{sat}}}{2}\right)V_{D\text{sat}}. \tag{36}$$

Equating Eqs. 36 and 35, we can solve for $V_{D\text{sat}}$ as

$$V_{D\text{sat}} = L\mathscr{E}_c + \frac{V_G - V_T}{M} - \sqrt{(L\mathscr{E}_c)^2 + \left(\frac{V_G - V_T}{M}\right)^2}. \tag{37}$$

Since $V_{D\text{sat}}$ here is always smaller than $(V_G - V_T)/M$ (Eq. 29), the field-dependent mobility always gives a lower $I_{D\text{sat}}$.

Field-Dependent Mobility: Empirical Formula. Next we consider the v-$\mathscr{E}$ relationship of Eq. 33 with $n = 1$. Substituting this into Eq. 21 gives

$$I_D\left(\mathscr{E}_c + \frac{d\Delta\psi_i}{dy}\right) = ZC_{ox}\mu_n\mathscr{E}_c(V_G - V_T - M\Delta\psi_i)\frac{d\Delta\psi_i}{dy}. \tag{38}$$

Notice that the right-hand side of the equation is similar to the constant-mobility model. Integrating the above equation from source to drain gives

$$I_D = \frac{ZC_{ox}\mu_n\mathscr{E}_c}{L\mathscr{E}_c + V_D}\left(V_G - V_T - \frac{MV_D}{2}\right)V_D. \tag{39}$$

This equation is similar to Eq. 31 when L is replaced with $L + V_D/\mathscr{E}_c$. Furthermore, $V_{D\text{sat}}$ is obtained by setting $dI_D/dV_D = 0$,

$$V_{D\text{sat}} = L\mathscr{E}_c\left[\sqrt{1 + \frac{2(V_G - V_T)}{ML\mathscr{E}_c}} - 1\right]. \tag{40}$$

Again, once $V_{D\text{sat}}$ is known, $I_{D\text{sat}}$ can be calculated from Eq. 39.

Velocity Saturation. Using either assumption described above, it is interesting and insightful to look at the extreme case of short-channel devices where velocity saturation completely limits the current flow. In such case we set $v = v_s$, and consequently Q_n has to be fixed for current continuity, and is approximated to be $(V_G - V_T)C_{ox}$. Equation 22 then becomes

$$I_{D\text{sat}} = \frac{Z}{L}\int_0^L |Q_n(y)|v(y)dy = \frac{Z}{L}|Q_n|v_s L$$
$$= Z(V_G - V_T)C_{ox}v_s \ . \tag{41}$$

The transconductance becomes

$$g_m \equiv \frac{dI_{D\text{sat}}}{dV_G} = ZC_{ox}v_s \tag{42}$$

and it is independent of gate bias.

To compare models based on constant mobility and velocity saturation, we show *I-V* curves of identical devices in Fig. 12. Several observations can be made. First, $I_{D\text{sat}}$ and $V_{D\text{sat}}$ are both lowered by velocity saturation, while the linear regions remain similar. The g_m (which is the current difference between V_G steps) also becomes a constant, independent of V_G Finally Eq. 41 shows an interesting phenomena that the saturation current no longer depends on the channel length.

Experimental data confirm that such simple theory is quite satisfactory. In reality, as Fig. 11 indicates, the carrier velocity never reaches exactly v_s. Also the lateral field is not uniform throughout the channel. It is more difficult for the lower field near the source to reach $\mathscr{E}_c$ and that presents a bottle-neck for the maximum current. A better agreement can often be reached by adding a pre-factor with a value of ≈ 0.5 – 1.0 for Eqs. 41 and 42.

Ballistic Transport. In the above section, velocity saturation is a steady-state, equilibrium phenomena at high field, when many scattering events are allowed to happen. However, in ultra-short channel lengths whose dimensions are on the order of or shorter than the mean free path, channel carriers do not suffer from scattering. They can gain energy from the field without losing it to the lattice through scattering, and can acquire a velocity much higher than the saturation velocity. This effect is called

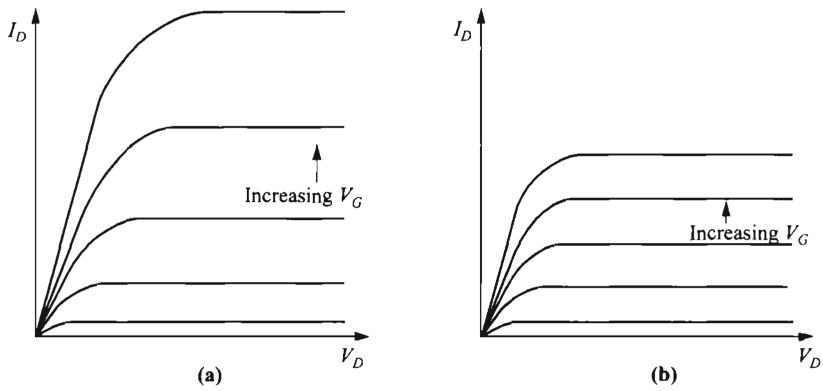


Fig. 12 Comparison of *I-V* characteristics for (a) constant mobility and (b) velocity saturation. All other parameters are the same.

ballistic transport (or velocity overshoot by some), introduced in Section 1.5.3. The ballistic transport is important since it points out that the current and transconductance can be higher than that of saturation velocity, giving an additional incentive for shrinking the channel length. The basic theory and insight in this topic are presented in Refs. 24–28.

Computer simulations show that in these devices, the field and velocity are very nonuniform. These quantities are qualitatively shown in Fig. 13. Notice that the longitudinal field along the channel (dE_C/dy) varies monotonically, being the highest at the drain end. Ballisticity always starts at the drain end, where the velocity can exceed the value of saturation velocity v_s. (For silicon at room temperature, $v_s \approx v_{th}$, the thermal velocity.) At positions closer to the source, the velocity decreases. In order to have current continuity, the channel potential and inversion charge must adjust themselves such that the product of velocity and charge would remain constant throughout the channel. By such argument, the bottleneck for the current flow, at the extreme of ultra-short channel length, would be at the position of maximum charge and minimum field, which means the potential maximum near the source end, indicated in Fig. 13a.

In analyzing the saturation current in the ballistic regime, we go back to the general equation of Eq. 22, and apply it to this maximum-potential point. We start with the generalized form

$$I_{D\text{sat}} = Z|Q_n|v_{eff} \tag{43}$$

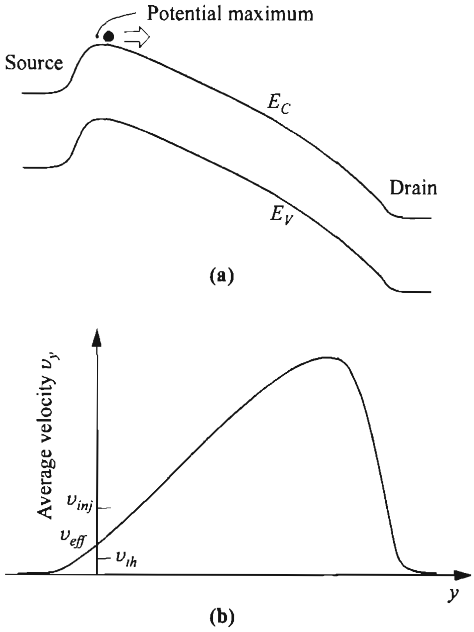


Fig. 13 (a) Under a drain bias, the potential maximum is the bottleneck for the current flow and is used to calculate the current. (b) Average carrier velocity (y-component) as a function of the channel position. Note that v_{eff} is between the values of v_{inj} and v_{th}, and that the maximum velocity near the drain can be higher than v_{inj}.

where $|Q_n|$ has the maximum value at the source as $C_{ox}(V_G - V_T)$, and v_{eff} is an effective average carrier velocity that should match the final experimental saturation current. So in this simple expression, the only critical parameter is v_{eff}.

The maximum value of v_{eff}, according to classical thermal equilibrium, is simply the thermal velocity v_{th} $[= (2kT/\pi m^*)^{1/2}]$. Close examination of the system reveals that for higher inversion charge density, the random velocity can exceed this thermal limit. This is a quantum-mechanical effect, called carrier degeneracy,[25] where the mean carrier energy is pushed to a higher state than the thermal energy. This higher value is called the injection velocity v_{inj}, related to the Fermi energy with respect to the quantized energy E_n inside the potential well where carriers reside, given by[24]

$$v_{inj} = \sqrt{\frac{2kT}{\pi m^*}} \frac{F_{1/2}[(E_F - E_n)/kT]}{\ln\{1 + \exp[(E_F - E_n)/kT]\}}, \tag{44}$$

where $F_{1/2}$ is the Fermi-Dirac integral (see Section 1.4.1). With a small inversion charge or $E_F - E_n$, Eq. 44 reduces to $\sqrt{2kT/\pi m^*}$, and $v_{inj} = v_{th}$. If the inversion charge is high, Eq. 44 can be simplified to

$$v_{inj} = \frac{8\hbar}{3m^*}\sqrt{\frac{|Q_n|}{2\pi q}} = \frac{8\hbar}{3m^*}\sqrt{\frac{C_{ox}(V_G - V_T)}{2\pi q}}, \tag{45}$$

and is a function of the inversion charge or gate overdrive. Theoretical v_{inj} as a function of the inversion charge is shown in Fig. 14. The maximum current, which is a product of $Q_n v_{inj}$, gives the ultimate current drive of the ballistic MOSFET and is also plotted in the same figure.

The saturation current of Eq. 43 can be rewritten as

$$\begin{aligned} I_{Dsat} &= r_n Z |Q_n| v_{inj} \\ &= \frac{8 r_n Z \hbar}{3m^*} \frac{[C_{ox}(V_G - V_T)]^{3/2}}{\sqrt{2\pi q}} \end{aligned} \tag{46}$$

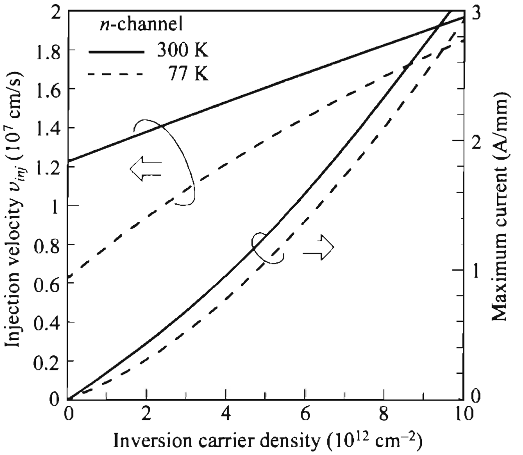


Fig. 14 Injection velocity v_{inj} as a function of inversion density. The product of v_{inj} and inversion charge gives the maximum current. (After Ref. 24.)

where r_n is the *index of ballisticity* ($= v_{eff}/v_{inj}$). In the extreme of ballistic transport, $r_n = 1$, and it sets the ultimate current drive for $L \to 0$. The transconductance is given by

$$g_m = \frac{4r_n Z\hbar}{m^*}\sqrt{\frac{C_{ox}(V_G - V_T)}{2\pi q}}. \tag{47}$$

It is seen here that both $I_{D\text{sat}}$ and g_m are independent of channel length L.

The index of ballisticity is also interpreted by back scattering R of channel carriers at the drain back to the source. Furthermore, since mobility is also a consequence of scattering, there should be some relationship between r_n and the low-field mobility μ_n. It has been shown that:[26]

$$\begin{aligned} r_n &= \frac{v_{eff}}{v_{inj}} = \frac{1-R}{1+R} \\ &= \left[\frac{1}{v_{inj}} + \frac{1}{\mu_n \mathscr{E}(0^+)}\right]^{-1}, \end{aligned} \tag{48}$$

where $\mathscr{E}(0^+)$ is the field at a potential kT down from the maximum toward the drain. With this interpretation, some experimental data and simulation trends are nicely explained. It has been seen that at lower temperature, $I_{D\text{sat}}$ is increased, and that at the same temperature, higher low-field mobility always gives higher $I_{D\text{sat}}$ even in the ballistic regime. Both of these can be explained through improvement of μ_n in Eq. 48.

It should be emphasized that in this model, at or near the maximum-potential point, the field is too low to cause ballistic transport, so v_{inj} sets the maximum current, even though a location near the drain can have ballistic transport. The high ballistic velocities near the drain cannot produce a higher current than v_{inj} can support, but it helps to achieve this maximum value set by v_{inj} by rebalancing the whole system.

It is interesting to compare the V_G dependence of $I_{D\text{sat}}$ for different channel lengths. In the long-channel, constant-mobility regime, $I_{D\text{sat}} \propto (V_G - V_T)^2$. In the short-channel, saturation-velocity regime, $I_{D\text{sat}} \propto (V_G - V_T)$. And in the limit of ballistic regime, $I_{D\text{sat}} \propto (V_G - V_T)^{3/2}$.

6.2.3 Threshold Voltage

We now return to the discussion of threshold voltage, first mentioned in Eq. 25. To account for the threshold shift from nonzero flat-band voltage whose main cause comes from fixed oxide charges Q_f and the work-function difference ϕ_{ms} between the gate material and the semiconductor, Eq. 25 becomes

$$\begin{aligned} V_T &= V_{FB} + 2\psi_B + \frac{\sqrt{2\varepsilon_s q N_A (2\psi_B)}}{C_{ox}} \\ &= \left(\phi_{ms} - \frac{Q_f}{C_{ox}}\right) + 2\psi_B + \frac{\sqrt{4\varepsilon_s q N_A \psi_B}}{C_{ox}}. \end{aligned} \tag{49}$$

Qualitatively, V_T is the gate bias beyond flat-band just starting to induce an inversion charge sheet and is given by the sum of voltages across the semiconductor ($2\psi_B$) and the oxide layer (last term of Eq. 25). The square-root term is the total depletion-layer charge.

When a substrate bias is applied (negative for n-channel or p-substrate), the threshold voltage becomes

$$V_T = V_{FB} + 2\psi_B + \frac{\sqrt{2\varepsilon_s q N_A (2\psi_B - V_{BS})}}{C_{ox}}, \tag{50}$$

and it is shifted by the amount of

$$\Delta V_T = V_T(V_{BS}) - V_T(V_{BS}=0) = \frac{\sqrt{2\varepsilon_s q N_A}}{C_{ox}}\left(\sqrt{2\psi_B - V_{BS}} - \sqrt{2\psi_B}\right). \tag{51}$$

In practice it is often necessary to minimize this threshold-voltage shift due to substrate bias. In these cases, low substrate doping and thin oxide thickness are preferred.

To measure the threshold voltage, we use the linear region by applying a small drain bias ($V_D \ll V_G$), and plot I_D versus V_G as shown in Fig. 15a. According to Eq. 24, the extrapolated value at the V_G axis is equal to $V_T + \frac{1}{2}V_D$. Below the threshold voltage, I_D is considered zero in the linear scale, but details can be displayed in the logarithmic scale (Fig. 15b).

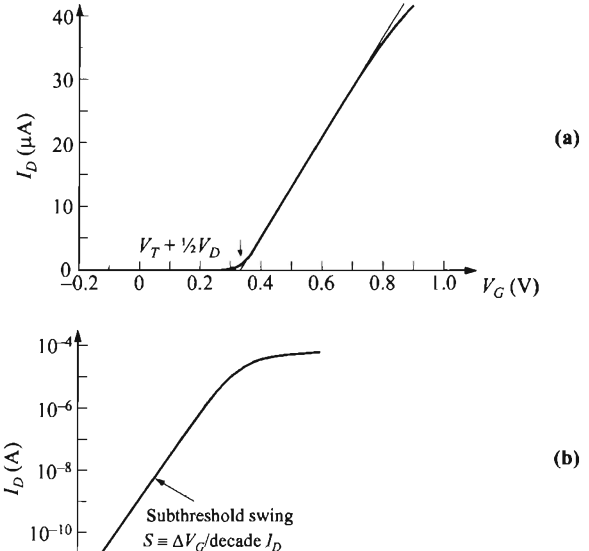


Fig. 15 Transfer characteristics (I_D vs. V_G) in the linear region ($V_D \ll V_G$). (a) I_D in linear scale to deduce V_T. Deviation from linearity at higher V_G is due to lower mobility. (b) I_D in logarithmic scale to show subthreshold swing.

6.2.4 Subthreshold Region

When the gate bias is below the threshold and the semiconductor surface is in weak inversion or depletion, the corresponding drain current is called the subthreshold current.[29,30] The subthreshold region tells how sharply the current drops with gate bias and is particularly important for low-voltage, low-power applications, such as when the MOSFET is used as a switch in digital logic and memory applications.

In weak inversion and depletion, the electron charge is small and, thus, the drift current is low. The drain current is dominated by diffusion and is derived in the same way as the collector current in a bipolar transistor with homogeneous base doping. Considering the electron-density gradient in the channel, the diffusion current is given by

$$I_D = -ZqD_n\frac{dN'(y)}{dy} \approx ZqD_n\frac{N'(0)-N'(L)}{L}, \tag{52}$$

where N' is the electron density per unit area, integrated over the depletion width. The electron density at the source end is given by

$$N'(0) = \int_0^{W_D} n(x)\,dx = n_{po}\int_{\psi_s}^{0} \exp(\beta\psi_p)\,d\psi_p. \tag{53}$$

Since the potential distribution inside the depletion region is known, this electron density can be calculated to be[31]

$$N'(0) \approx \left(\frac{1}{\beta}\right)\sqrt{\frac{\varepsilon_s}{2q\,\psi_s N_A}}\;n_{po}\exp(\beta\psi_s). \tag{54}$$

Similar result can be obtained by assuming an effective thickness (x_i) of the surface charge layer. Because of the exponential dependence of electron density on the potential ψ_p, x_i corresponds to the distance in which ψ_p decreases by kT/q. Therefore, x_i is $kT/q\mathscr{E}_s$ where $\mathscr{E}_s$ is the semiconductor surface field. With this assumption, we get the same expression

$$\begin{aligned} N'(0) &= x_i \times n(x{=}0) = \left(\frac{kT}{q\mathscr{E}_s}\right) n_{po}\exp(\beta\psi_s) \\ &= \left(\frac{kT}{q}\right)\sqrt{\frac{\varepsilon_s}{2q\,\psi_s N_A}}\;n_{po}\exp(\beta\psi_s)\;. \end{aligned} \tag{55}$$

The electron density at the drain end is lowered exponentially by the drain bias,

$$N'(L) = N'(0)\exp(-\beta V_D). \tag{56}$$

Substituting Eqs. 55 and 56 into Eq. 52 gives

$$\begin{aligned} I_D &= \frac{Z\mu_n}{L\beta^2}\sqrt{\frac{q\varepsilon_s N_A}{2\psi_s}}\left(\frac{n_i}{N_A}\right)^2 \exp(\beta\psi_s)[1-\exp(-\beta V_D)] \\ &\approx \frac{Z\mu_n}{L\beta^2}\sqrt{\frac{q\varepsilon_s N_A}{2\psi_s}}\left(\frac{n_i}{N_A}\right)^2 \exp(\beta\psi_s) \end{aligned} \tag{57}$$

when $V_D \gg kT/q$. Equation 57 indicates that in the subthreshold region the drain current varies exponentially with ψ_s, and for drain voltage V_D larger than $\approx 3kT/q$, the current becomes independent of V_D. Next, in order to relate current to the gate bias, the relationship between V_G and ψ_s is needed.

From Chapter 4 on the MOS capacitor, we have the following relationship (Eq. 33):

$$V_G - V_{FB} = \psi_s + \frac{\sqrt{2\varepsilon_s \psi_s q N_A}}{C_{ox}}. \tag{58}$$

This quadratic equation will not give a simple expression of ψ_s as a function of V_G. But once ψ_s is known from Eq. 58, the subthreshold current can be calculated.

The parameter to quantify how sharply the transistor is turned off by the gate voltage is called the subthreshold swing S (inverse of subthreshold slope), defined as the gate-voltage change needed to induce a drain-current change of one order of magnitude. First, from Eq. 58, the relative change of V_G and ψ_s is calculated to be

$$\frac{dV_G}{d\psi_s} = 1 + \frac{1}{C_{ox}}\sqrt{\frac{\varepsilon_s q N_A}{2\psi_s}} = \frac{C_{ox} + C_D}{C_{ox}}. \tag{59}$$

By definition, the subthreshold swing can now be calculated:

$$\begin{aligned} S &\equiv (\ln 10)\frac{dV_G}{d(\ln I_D)} = (\ln 10)\frac{dV_G}{d(\beta\psi_s)} \\ &= (\ln 10)\left(\frac{kT}{q}\right)\left(\frac{C_{ox} + C_D}{C_{ox}}\right) . \end{aligned} \tag{60}$$

Note that ψ_s in the square-root term in Eq. 57 is treated as a constant since it is a much weaker function compared to the exponential term.

Having derived the subthreshold swing in Eq. 60, it is intuitive to explain its simple form. In the extreme of zero oxide thickness, the exponential characteristics are identical to the familiar case of the diffusion current in a *p-n* junction. For nonzero oxide thickness, the swing is just degraded by a factor which is a voltage divider of two capacitors in series, whose ratio is $(C_{ox} + C_D)/C_{ox}$. The voltage divider is exactly the implication of Eq. 59. One also notices that since the depletion width (and C_D) varies with ψ_s, the subthreshold swing is a weak function, but not exactly constant with V_G.

In the presence of a significant interface-trap density D_{it}, its associated capacitance C_{it} $(= q^2 D_{it})$ is in parallel with the depletion-layer capacitance C_D. Using Eq. 60 and substituting $(C_D + C_{it})$ for C_D (see Fig. 14 of Chapter 4), we obtain

$$\begin{aligned} S(\text{with } D_{it}) &= (\ln 10)\left(\frac{kT}{q}\right)\left(\frac{C_{ox} + C_D + C_{it}}{C_{ox}}\right) \\ &= S(\text{without } D_{it}) \times \frac{C_{ox} + C_D + C_{it}}{C_{ox} + C_D}. \end{aligned} \tag{61}$$

If other device parameters such as doping and oxide thickness are known, by measuring the subthreshold swing, the interface-trap density can be obtained. This provides an attractive option in measuring D_{it} besides using the MOS capacitor in which ac measurements have to be made. In general, dc I-V measurements are much easier to make than ac capacitance and conductance, provided a three-terminal transistor structure is available (substrate contact is not critical here).

For a sharp subthreshold slope (small S), it is preferable to have low channel doping, thin oxide thickness, low interface-trap density, and low-temperature operation. When a substrate bias is applied, in addition to shifting the threshold voltage, it increases the value of ψ_s by V_{BS}. Consequently, the depletion-layer capacitance C_D is reduced and therefore S is reduced.

In Fig. 15a, at and near the threshold voltage, the drain current does not turn off as sharply as Eq. 24 predicts. This is due to diffusion current which is the dominant current near and below threshold, and it has been ignored so far, as one of the assumptions made at the beginning of Section 6.2.2. To consider the effect of the diffusion component, we refer to Fig. 7 for the nonequilibrium condition. The total drain current density including both drift and diffusion components is given by

$$J_D(x,y) = q\mu_n n\mathscr{E}_y + qD_n\frac{dn}{dy} = D_n n(x,y)\frac{dE_{Fn}}{dy}. \tag{62}$$

The drain current based on the gradual-channel approximation is

$$\begin{aligned} I_D &= Z\int_0^{x_i} J_D(x,y)dx = \frac{ZD_n}{L}\int_0^L \frac{dE_{Fn}}{dy}\int_0^{x_i} n(x,y)dxdy \\ &= \frac{Z\varepsilon_s\mu_n}{L\,L_D}\int_0^{V_D}\int_{\psi_B}^{\psi_s}\frac{\exp(\beta\psi_p - \beta\Delta\psi_i)}{F(\beta\psi_p, \Delta\psi_i, n_{po}/p_{po})}d\psi_p d\Delta\psi_i\ . \end{aligned} \tag{63}$$

The gate voltage V_G is related to the surface potential ψ_s by

$$\begin{aligned} V_G - V_{FB} &= -\frac{Q_s}{C_{ox}} + \psi_s \\ &= \frac{2\varepsilon_s kT}{C_{ox}qL_D}F\left(\beta\psi_s, \Delta\psi_i, \frac{n_{po}}{p_{po}}\right) + \psi_s\ . \end{aligned} \tag{64}$$

Equation 63 reduces to Eq. 23 for gate voltages well above threshold. The latter, however, becomes inaccurate for gate voltages near and below threshold and near the pinch-off point. For a particular device with known physical dimensions and other device parameters, Eq. 63 can be calculated numerically to give accurate results for the entire range of drain voltage, from the linear region to the saturation region.

6.2.5 Mobility Behavior

Because channel carriers are confined to a thin inversion layer, their drift velocity v and mobility μ are expected to be influenced by the thickness of this inversion layer. When a small longitudinal field $\mathscr{E}_y$ is applied (parallel to the semiconductor surface), the drift velocity varies linearly with $\mathscr{E}_y$ and the proportionality constant is the low-field mobility. Experimental measurements on Si inversion layers show that this low-field mobility, while independent of $\mathscr{E}_y$, is a unique function of the transverse field $\mathscr{E}_x$ that is perpendicular to the current flow.[32] This dependence is not directly on the oxide thickness or doping density, but through their impact of $\mathscr{E}_x$ in the inversion layer. The measured results are shown in Fig. 16. When many devices with different oxide thicknesses and doping levels are measured, the mobility is found to correlate well with a single parameter that is related to $\mathscr{E}_x$. At a given temperature, mobility decreases with an increasing *effective* transverse field, defined as the field averaged over the electron distribution in the inversion layer, given by

$$(\mathscr{E}_x)_{\text{eff}} = \frac{1}{\varepsilon_s}\left(Q_B + \frac{1}{2}Q_n\right). \tag{65}$$

Physically it means an average inversion carrier experiences the full effect of the depletion-layer charge Q_B, but only half of the inversion-layer charge Q_n. Note that this effective mobility is valid for the current expressions in Eqs. 24 and 27, but will be slightly different from that in the g_m expression (such as Eq. 30) in which constant mobility has been assumed.

When the longitudinal field increases, the υ-$\mathscr{E}$ relationship starts to deviate from linearity. This field-dependent mobility has been discussed and generally described by Eq. 33 and shown in Fig, 11. Measured electron drift velocity as a function of $\mathscr{E}_y$ for various $\mathscr{E}_x$ is shown in Fig. 17. Since mobility at any field is defined as the ratio of $\upsilon/\mathscr{E}_y$, it decreases monotonically with $\mathscr{E}_y$. Eventually velocity saturation occurs and results in a value similar to that of bulk silicon. The influence of $\mathscr{E}_x$ on low-field

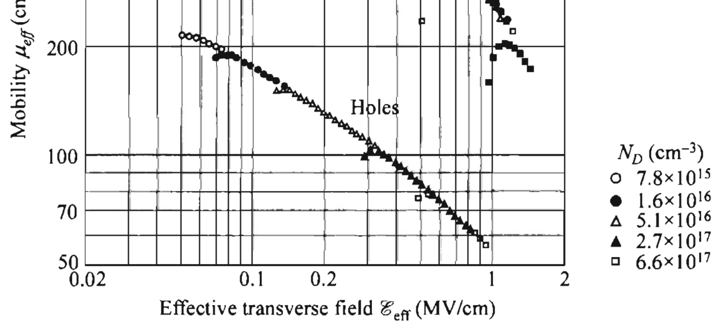


Fig. 16 Electron and hole inversion-layer mobilities vs. effective transverse field, at room temperature on Si (100) surface. (After Ref. 33.)

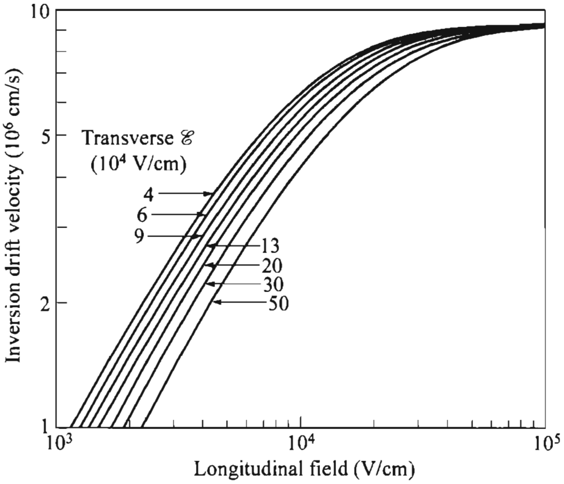


Fig. 17 Electron surface drift velocity vs. longitudinal field for various transverse fields. The slope at low longitudinal field is mobility. (After Ref. 34.)

mobility from Fig. 16 is also reflected in this figure. It can be seen here also that the saturation velocity v_s is independent of the low-field mobility or $\mathscr{E}_x$.

6.2.6 Temperature Dependence

Temperature affects device parameters and performance, in particular mobility, threshold voltage, and subthreshold characteristics. The effective mobility in inversion layer has a T^{-2} power dependence on temperatures around 300 K at gate biases corresponding to strong inversion.[32] This gives rise to higher current and transconductance at lower temperature.

To derive the temperature dependence of the threshold voltage, we repeat the expression from Eq. 49:

$$V_T = \phi_{ms} - \frac{Q_f}{C_{ox}} + 2\psi_B + \frac{\sqrt{4\varepsilon_s q N_A \psi_B}}{C_{ox}}. \tag{66}$$

Because the work-function difference ϕ_{ms} and the fixed oxide charges are essentially independent of temperature, differentiating Eq. 66 with respect to temperature yields[35]

$$\frac{dV_T}{dT} = \frac{d\psi_B}{dT}\left(2 + \frac{1}{C_{ox}}\sqrt{\frac{\varepsilon_s q N_A}{\psi_B}}\right). \tag{67}$$

From the basic equations of

$$\psi_B = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right), \tag{68}$$

$$n_i^2 \propto T^3 \exp\left(\frac{-E_{g0}}{kT}\right), \tag{69}$$

where E_{g0} is the energy gap at $T = 0$, we obtain

$$\frac{d\psi_B}{dT} \approx \frac{1}{T}\left(\psi_B - \frac{E_{g0}}{2q}\right). \tag{70}$$

Figure 18 shows the results of such calculations at room temperature as a function of substrate doping for various values of oxide thickness. Note that depending on the oxide thickness, the quantity $|dV_T/dT|$ can increase or increase with the substrate doping.

As temperature decreases, the MOSFET characteristics improve, especially in the subthreshold region. Figure 19 shows the transfer characteristics of a long-channel MOSFET ($L = 9$ µm) with temperature as a parameter. Note that as temperature decreases from 296 K to 77 K, the threshold voltage V_T increases from 0.25 V to about 0.5 V. This increase in V_T is similar to that shown in Fig. 18. The most-important improvement is the reduction of the subthreshold swing S, from 80 mV/decade at 296 K to 22 mV/decade at 77 K. Thus, the improvement in the subthreshold swing at 77 K is about a factor of four. This improvement comes mainly from the kT/q term in Eq. 60. Other improvements at 77 K include higher mobility, thus, higher current and transconductance, lower power consumption, lower junction leakage current, and lower metal-line resistance. The major disadvantages are that the MOSFET must be immersed in a suitable inert coolant (e.g., liquid nitrogen), and that a low-temperature setup requires additional equipment and special care.

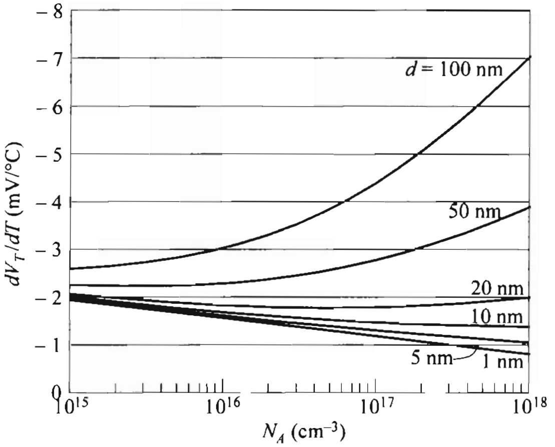


Fig. 18 Threshold-voltage shift (dV_T/dT) of a Si-SiO$_2$ system at room temperature vs. substrate doping, with oxide thickness d as a parameter.

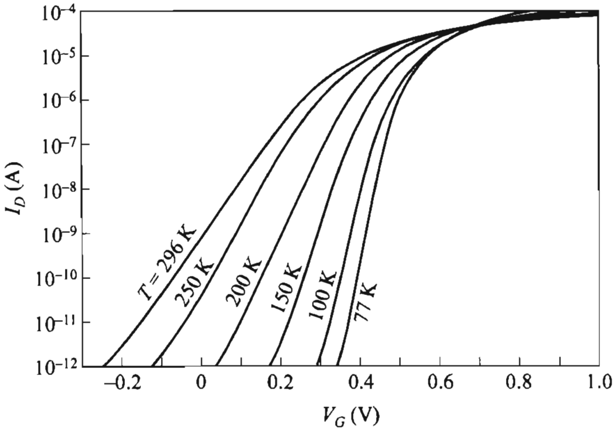


Fig. 19 Subthreshold characteristics for a long-channel MOSFET ($L = 9$ μm) with temperature as a parameter. (After Ref. 36.)

6.3 NONUNIFORM DOPING AND BURIED-CHANNEL DEVICE

In Section 6.2 doping concentration in the channel is assumed to be constant. In practical devices, however, the doping is generally nonuniform because in modern MOSFET technology, ion implantation is used extensively to taylor the doping profile and improve the device performance for specific applications. For example, a lighter doping at deeper region reduces drain-substrate capacitance and also the substrate-bias effect. On the other hand, a lighter doping level near the Si-SiO$_2$ interface lowers the threshold voltage, reduces the field and improves mobility, and higher level at deeper region reduces punch-through between source and drain. These two general cases, namely *high-low* and *low-high* profiles, are depicted in Fig. 20, with their step-profile approximations for ease of analysis.

We consider next the effect of nonuniform channel doping on device characteristics, especially on threshold voltage and depletion width which in turn affects subthreshold swing and the substrate-bias effect. Note that what is most important for determining V_T is the doping profile within the depletion region. The profile outside the depletion is important for considerations of capacitance and substrate sensitivity, that is, dependence of the threshold voltage on the substrate reverse bias. With that in mind, the general equation for the threshold voltage is given by

$$\begin{aligned} V_T &= V_{FB} + \psi_s + \frac{Q_B}{C_{ox}} \\ &= V_{FB} + 2\psi_B + \frac{q}{C_{ox}}\int_0^{W_{Dm}} N(x)dx \, , \end{aligned} \tag{71}$$

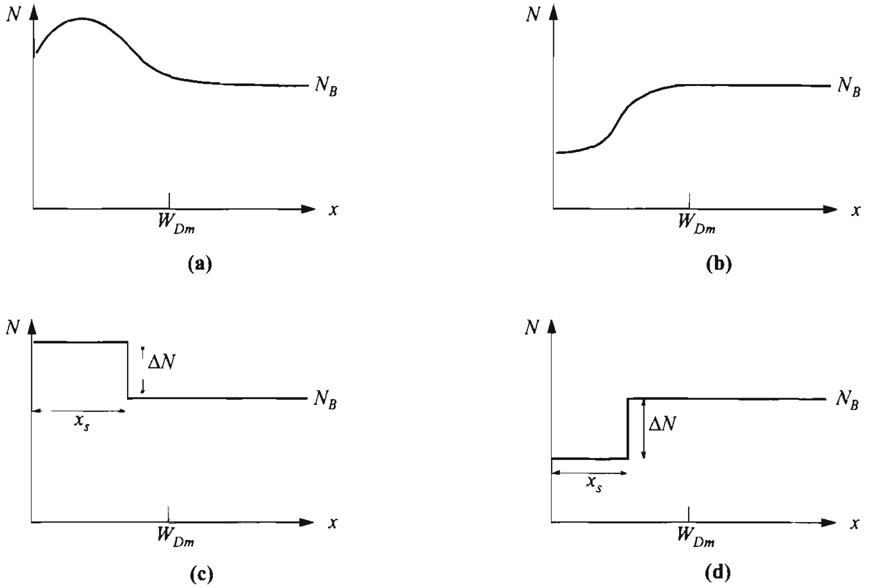


Fig. 20 Nonuniform channel doping profiles. (a) High-low profile. (b) Low-high (retrograde) profile. (c) – (d) Their approximations using step profiles.

where Q_B is the depletion-layer charge. The limit for integration, that is the maximum depletion width W_{Dm}, is needed and determined by the Poisson equation with the onset of strong inversion being the boundary condition,

$$\psi_s = 2\psi_B = \frac{q}{\varepsilon_s}\int_0^{W_{Dm}} xN(x)dx. \tag{72}$$

Note that for a nonuniform profile, the definitions of ψ_B and V_{FB} become nontrivial and complicated. Fortunately, using the background doping of N_B for these values is found to be sufficiently accurate. This is especially true for the surface potential $\psi_s = 2\psi_B$ since it is a very weak function of doping level.

6.3.1 High-Low Profile

To derive the threshold voltage shift due to ion implantation, we shall consider an idealized step profile as shown in Fig. 20c. The implant profile, after thermal anneal, is approximated by the step function with step depth x_s, roughly equal to the sum of the projected range and the standard deviation of the original implant. For a wider x_s, that is, the maximum depletion-layer width W_{Dm} under heavy inversion is within x_s, the surface region can be considered a uniformly doped region with a higher concentration. The threshold voltage is identical to that given by Eq. 50. If $W_{Dm} > x_s$, the threshold voltage is obtained from Eq. 71,

$$
\begin{aligned}
V_T &= V_{FB} + 2\psi_B + \frac{qN_B W_{Dm} + q\Delta N x_s}{C_{ox}} \\
&= V_{FB} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2q\varepsilon_s N_B\left(2\psi_B - \frac{q\Delta N x_s^2}{2\varepsilon_s}\right)} + \frac{q\Delta N x_s}{C_{ox}}.
\end{aligned}
\tag{73}
$$

The depletion width can be obtained from Eq. 72, using $\psi_s = 2\psi_B$ for strong inversion:

$$
W_{Dm} = \sqrt{\frac{2\varepsilon_s}{qN_B}\left(2\psi_B - \frac{q\Delta N x_s^2}{2\varepsilon_s}\right)}. \tag{74}
$$

From these equations, we see that added surface doping increases V_T and decreases W_{Dm}.

Notice that for the same dose, the V_T shift is largest with the added doping closest to the surface. For the limiting case of a delta function of dose localized at the Si-SiO$_2$ interface ($x_s = 0$), the threshold shift is simply

$$
\Delta V_T \approx \frac{qD_I}{C_{ox}}, \tag{75}
$$

where D_I is the total dose $\Delta N x_s$. Such approach is called threshold adjust, which has the same effect as changing the work-function difference ϕ_{ms} or changing the total fixed oxide charge.

The step-profile approach described above can give first-order results for the threshold voltage. To obtain a more accurate V_T, we have to consider the actual doping profile, because the step width x_s is not well defined for nonuniform doping. A schematic diagram for the nonuniform implanted doping $N(x)$ is shown in Fig. 21. For a typical case, the threshold voltage depends on the implanted dose D_I and the centroid of the dose x_c. Therefore, the actual implant can be replaced by a delta-function located at $x = x_c$ as shown;

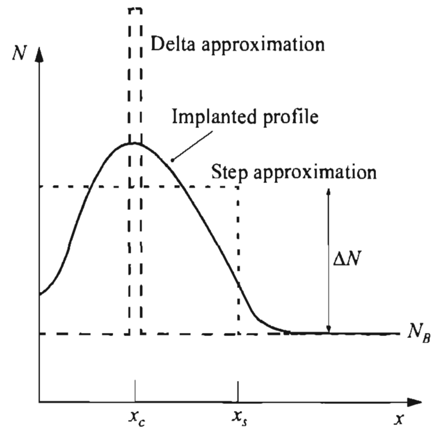


Fig. 21 Approximation of an actual implanted profile by step and delta profiles.

$$D_I = \int_0^{W_{Dm}} \Delta N(x)dx, \tag{76}$$

$$x_c = \frac{1}{D_I}\int_0^{W_{Dm}} x\Delta N(x)dx. \tag{77}$$

With these, Eqs. 73 and 74 are now generalized to:

$$V_T = V_{FB} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2q\varepsilon_s N_B\left(2\psi_B - \frac{qx_cD_I}{\varepsilon_s}\right)} + \frac{qD_I}{C_{ox}}, \tag{78}$$

$$W_{Dm} = \sqrt{\frac{2\varepsilon_s}{qN_B}\left(2\psi_B - \frac{qD_Ix_c}{\varepsilon_s}\right)}. \tag{79}$$

It is interesting to examine the dependence of the threshold voltage shift and depletion width on the centroid x_c for a given dose D_I. For $x_c = 0$, the implant is a delta function at the Si-SiO$_2$ interface, and Eq. 78 gives $\Delta V_T = qD_I/C_{ox}$, which is the same as Eq, 75. As x_c increases, the dose becomes less effective in changing V_T, and the depletion width W_{Dm} decreases also at the same time. Eventually x_c meets the depletion edge, and then W_{Dm} becomes clamped to and increases with the implant centroid x_c. The condition for which x_c starts to be equal to W_{Dm} can be obtained from Eq. 79:

$$D_I(x_c = W_{Dm}) = \frac{N_B(W_{Dm0}^2 - x_c^2)}{2x_c}, \tag{80}$$

where W_{Dm0} is the original W_{Dm} with background doping N_B. Eventually, as x_c moves beyond the W_{Dm0}, it no longer has any effect on threshold voltage and depletion width.

To consider the subthreshold swing and the substrate sensitivity, in Section 6.2.4 we have interpreted the subthreshold swing by comparing the gate-oxide capacitance C_{ox} to depletion capacitance C_D. So, once the depletion width is known, the subthreshold swing can be calculated. For the high-low profile, the added doping decreases W_{Dm}, increases C_D, and results in a larger (less steep) subthreshold swing. The substrate sensitivity can be calculated also by substituting $2\psi_B$ with $2\psi_B + V_{BS}$ in calculating V_T.

6.3.2 Low-High Profile

Analysis of the low-high profile (Fig. 20b), also called the retrograde profile, is similar to the high-low case with a ΔN being subtracted from the background doping. The appropriate equations for the threshold voltage and depletion width become, with just a change of signs:

$$\begin{aligned} V_T &= V_{FB} + 2\psi_B + \frac{qN_B W_{Dm} - q\Delta N x_s}{C_{ox}} \\ &= V_{FB} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2q\varepsilon_s N_B\left(2\psi_B + \frac{q\Delta N x_s^2}{2\varepsilon_s}\right)} - \frac{q\Delta N x_s}{C_{ox}}, \end{aligned} \tag{81}$$

and

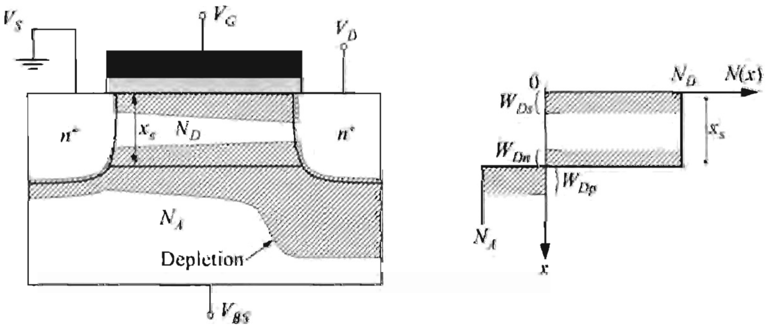
$$W_{Dm} = \sqrt{\frac{2\varepsilon_s}{qN_B}\left(2\psi_B + \frac{q\Delta N x_s^2}{2\varepsilon_s}\right)}. \tag{82}$$

The threshold voltage is, thus, decreased and the depletion width is increased by a dip at the surface doping.

6.3.3 Buried-Channel Device

In the extreme case of the low-high profile, the surface doping can be of the opposite type of the substrate. When this happens, and if part of the surface doped layer is not fully depleted, that is, there exists some neutral region, current can conduct through this buried layer. We call this type of device a buried-channel device.$^{37-40}$ Figure 22a shows a cross section of such a buried-n-channel MOSFET. The gate voltage can change the surface depletion layer, thus controlling the net opening of the channel thickness and controlling the current flow. With a large positive gate bias, the channel is fully open, and an addition surface inversion layer can be induced at the surface, similar to a regular surface channel, resulting in two channels in parallel.

The surface inversion channel has been the subject of discussion and needs no further elaboration. We now focus on the buried channel whose doping and dimensions are shown in Fig. 22b and whose energy-band diagrams are shown in Fig. 23. The net channel thickness is reduced from x_s by the amounts of surface depletion W_{Ds} and the bottom p-n junction depletion W_{Dn}. The surface depletion as a function of V_G is the same as Eq. 27 of Chapter 4, and is repeated and modified here as



(a) (b)

Fig. 22 (a) Schematic of a buried-channel MOSFET under bias. (b) Its doping profile and depletion regions.

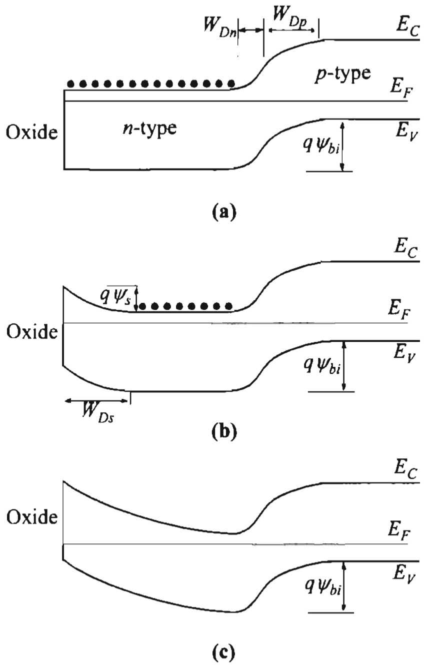


Fig. 23 Energy-band diagrams of a buried-channel MOSFET, for the bias conditions of (a) flat band ($V_G = V_{FB}^*$), (b) surface depletion, and (c) threshold ($V_G = V_T$).

$$W_{Ds} = \sqrt{\frac{2\varepsilon_s}{qN_D}(V_{FB}^* - V_G) + \frac{\varepsilon_s^2}{C_{ox}^2}} - \frac{\varepsilon_s}{C_{ox}} . \tag{83}$$

Note that the flat-band voltage V_{FB}^* takes on slightly different meaning here (Fig. 23a). We refer now to the condition that the surface n-layer has flat band, as opposed to the p-substrate. The new flat-band voltage is redefined as

$$V_{FB}^* = V_{FB} + \psi_{bi} \tag{84}$$

where V_{FB} keeps the p-substrate as reference. The bottom depletion width is from the p-n junction theory, given by

$$W_{Dn} = \sqrt{\frac{2\varepsilon_s \psi_{bi}}{qN_D}\left(\frac{N_A}{N_D + N_A}\right)} . \tag{85}$$

Of special interest is the threshold voltage V_T at which gate bias the channel width is totally consumed by both depletion regions. Setting the condition of

$$x_s = W_{Ds} + W_{Dn} , \tag{86}$$

the threshold voltage is obtained;[40]

$$V_T = V_{FB}^* - qN_D x_s\left(\frac{x_s}{2\varepsilon_s} + \frac{1}{C_{ox}}\right) + \left(\frac{x_s}{\varepsilon_s} + \frac{1}{C_{ox}}\right)\sqrt{\frac{2q\varepsilon_s N_D N_A \psi_{bi}}{N_D + N_A}} - \frac{N_A \psi_{bi}}{N_D + N_A} . \tag{87}$$

Once the channel dimensions are known, the channel charge can be calculated easily. Depending on the gate-bias range, we can have different amounts of bulk change Q_B and surface inversion charge Q_I. These are given as;

$$Q = Q_B = (x_s - W_{Ds} - W_{Dn})N_D \ , \qquad V_T < V_G < V^*_{FB} \ , \qquad (88)$$

and

$$\begin{aligned} Q &= Q_B + Q_I \\ &= (x_s - W_{Dn})N_D + C_{ox}(V_G - V^*_{FB}) \ , \qquad V^*_{FB} < V_G \ . \qquad (89) \end{aligned}$$

Given the channel charge, the drain current can be calculated in a way similar to those previously derived. But compared to the surface-channel devices, the buried-channel MOSFET equations are more complicated since the coupling of the gate to the channel (or net gate capacitance) is now gate-bias dependent. Qualitative I-V characteristics are shown in Fig. 24.

More-exact solution of the drain current can be obtained by substituting the charge into Eq. 22. The results, divided into different regimes of V_G bias, are summarized in Table 1. These results are based on the long-channel constant-mobility model. Current saturation due to velocity saturation can be estimated to be $\approx Qv_sW$.

A buried-channel MOSFET is usually normally-on (depletion-mode), although theoretically it can be made as a normally-off (enhancement-mode) device, by proper choice of metal work function, for example. Also for a given N_D, the threshold voltage becomes more negative with increased buried-channel depth x_s. However, because there exists a maximum depletion width in an MOS system, if the doping density N_D or/and the buried-channel depth x_s are sufficiently large, W_{Ds} can reach a maximum value without pinching off the channel. A limit on the channel profile thus exists, otherwise the transistor cannot be turned off. This condition is bound by a combination of x_s and N_D;

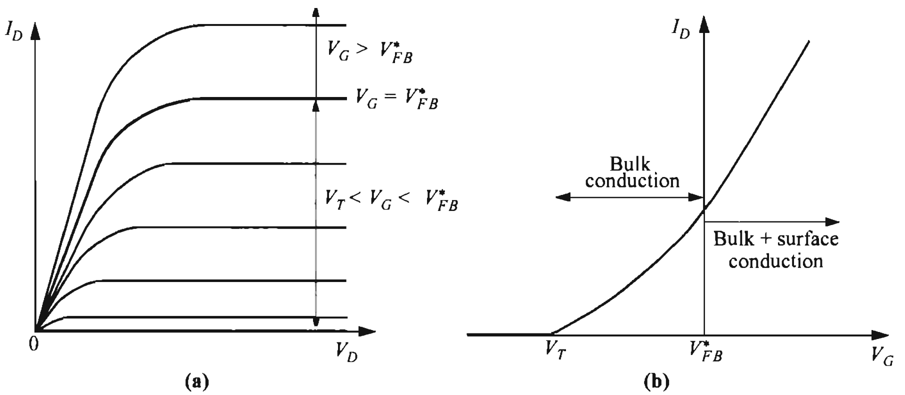


Fig. 24 Buried-channel MOSFET: (a) Output characteristics. (b) Transfer characteristics (I_D vs. V_G) in linear region (small V_D), showing threshold voltage V_T and flat-band voltage V^*_{FB} .

Table 1 Current Equations for Buried-Channel MOSFETs, Based on Long-Channel Constant Mobility (After Refs. 15 and 37)

$V_T \leq V_G \leq V_{FB}^*$

$$I_D = \frac{W}{L}\frac{\mu_B C_{ox}}{1+\sigma}\left[(V_G - V_T)V_D - \frac{1}{2}\alpha V_D^2\right], \qquad V_D \leq V_{D\text{sat}}$$

$$= \frac{W}{L}\frac{\mu_B C_{ox}}{1+\sigma}\frac{(V_G - V_T)^2}{2\alpha}. \qquad V_D \geq V_{D\text{sat}}$$

$V_G \geq V_{FB}^*$

$$I_D = \frac{W}{L}\frac{\mu_B C_{ox}}{1+\sigma}\left\{(V_G - V_T)V_D - \frac{1}{2}\alpha V_D^2 + (r-1)\left[(V_G - V_{FB}^*)V_D - \frac{1}{2}V_D^2\right]\right\}, \qquad V_D < V_G - V_{FB}^*$$

$$= \frac{W}{L}\frac{\mu_B C_{ox}}{1+\sigma}\left[(V_G - V_T)V_D - \frac{1}{2}\alpha V_D^2 + \frac{1}{2}(r-1)(V_G - V_{FB}^*)^2\right], \qquad V_G - V_{FB}^* \leq V_D < V_{D\text{sat}}$$

$$= \frac{W}{L}\frac{\mu_B C_{ox}}{1+\sigma}\left[\frac{(V_G - V_T)^2}{2\alpha} + \frac{1}{2}(r-1)(V_G - V_{FB}^*)^2\right]. \qquad V_D \geq V_{D\text{sat}}$$

where

$$V_{D\text{sat}} = (V_G - V_T)/\alpha \qquad \sigma = \frac{C_{ox}x_s}{\varepsilon_s}\left(\frac{C_{ox}x_s}{2\varepsilon_s} + 1\right) \qquad \alpha = 1 + (1+\sigma)\frac{\gamma}{4\sqrt{\psi_{bi}}}$$

μ_B = bulk mobility

μ_s = surface mobility

$$r = (1+\sigma)\frac{\mu_s}{\mu_B} \qquad \gamma = \frac{\sqrt{2\varepsilon_s q N_A}}{C_{ox}}$$

$$x_s\Big|_{\max} = \sqrt{\frac{2\varepsilon_s}{qN_D}}\left(\sqrt{2\psi_B} + \sqrt{\frac{N_A\psi_{bi}}{N_D + N_A}}\right). \tag{90}$$

In the buried-channel devices, the substrate-bias effect is more direct. It can be viewed as a bottom gate. The effects are calculated in the above equations if ψ_{bi} is replaced with $\psi_{bi} - V_{BS}$ (V_{BS} is negative). In particular, V_T (Eq. 87) and W_{Dn} (Eq. 85) can be shifted with a substrate bias, to the extent that the transistor can be turned on and off and between depletion-mode and enhancement-mode.

We now turn to the subthreshold current of buried-channel devices. At a sufficiently large negative gate bias, the channel will be pinched off, that is, when $x_x = W_{Ds} + W_{Dn}$ (Fig. 23c). The conduction below the threshold voltage is due to the presence of a region of partially depleted electrons, wherein the current is carried primarily by diffusion of electrons. The resulting subthreshold (sub-pinch-off) current for a buried-channel MOSFET is, thus, directly analogous to the subthreshold current for a surface-channel MOSFET. The subthreshold current will vary exponentially with the gate voltage, and the subthreshold swing S is given by the capacitive divider ratio again of Eq. 60, except now different capacitances have to be used. From Fig. 23c, the maximum electron concentration occurs at the location of $x \approx x_s - W_{Dn}$. So C_D of Eq. 60 should be replaced with the depletion capacitance of the substrate p-n junction

$[\varepsilon_s/(W_{Dn} + W_{Dp})]$, and C_{ox} should be replaced with C_{ox} in series with the surface depletion capacitance ε_s/W_{Ds}. These substitutions give an expression of[39]

$$S = (\ln 10)\frac{kT}{q}\left[1 + \frac{\varepsilon_{ox} W_{Ds} + \varepsilon_s d}{\varepsilon_{ox}(W_{Dn} + W_{Dp})}\right], \tag{91}$$

where all depletion layers W_{Ds}, W_{Dn}, and W_{Dp} correspond to the condition at threshold ($V_G = V_T$). The subthreshold swing is usually larger than that of conventional surface-channel devices.

The buried-channel device is expected to have higher carrier mobility than surface-channel devices since carriers are free of surface scattering and other surface effects. They are also less affected by the short-channel effects (to be discussed next) such as hot-carrier-induced reliability problems. On the other hand, since the net distance between the gate and the channel is further away and is gate-bias dependent, the transconductance is smaller and variable. Note that if the gate is replaced by a Schottky junction or a *p-n* junction, the device become a MESFET or a JFET correspondingly, both to be discussed in the next chapter.

6.4 DEVICE SCALING AND SHORT-CHANNEL EFFECTS

Since 1959, the beginning of the integrated-circuit era, the minimum feature length has been reduced by more than two orders of magnitude. We expect the minimum dimension will continue to shrink in the foreseeable future, as illustrated in Fig. 1. As the MOSFET dimensions shrink, they need to be designed properly to preserve the long-channel behavior as much as possible. As the channel length decreases, the depletion widths of the source and drain become comparable to the channel length and punch-through between the drain and source will eventually occur. This requires higher channel doping. A higher channel doping will increase the threshold voltage, and in order to control a reasonable threshold voltage, a thinner oxide is necessary. One sees that the device parameters are interrelated, and certain scaling rules are used to optimize the device performance.

Even with the best scaling rules, as the channel length is reduced, departures from long-channel behavior are inevitable. These departures, the short-channel effects, arise as results of a two-dimensional potential distribution and high electric fields in the channel region. The potential distribution in the channel now depends on both the transverse field $\mathscr{E}_x$ (controlled by the gate voltage and the back-substrate bias) and the longitudinal field $\mathscr{E}_y$ (controlled by the drain bias). In other words, the potential distribution becomes two-dimensional, and the gradual-channel approximation (that is, $\mathscr{E}_x \gg \mathscr{E}_y$) is no longer valid. This two-dimensional potential results in many forms of undesirable electrical behavior.

As the electric field is increased, the channel mobility becomes field-dependent, and eventually velocity saturation occurs. (The mobility behavior was discussed in Section 6.2.5.) When the field is increased further, carrier multiplication near the drain occurs, leading to substrate current and parasitic bipolar-transistor action. High

Table 2 MOSFET Scaling

Parameter	Scaling factor: Constant-$\mathscr{E}$	Scaling factor: Actual	Limitation
L	$1/\kappa$	/	/
$\mathscr{E}$	1	> 1	/
d	$1/\kappa$	$> 1/\kappa$	Tunneling, defects
r_j	$1/\kappa$	$> 1/\kappa$	Resistance
V_T	$1/\kappa$	$\gg 1/\kappa$	Off current
V_D	$1/\kappa$	$\gg 1/\kappa$	System, V_T
N_A	κ	$< \kappa$	Junction breakdown

In ideal constant-field scaling parameters are scaled by the same factor. In reality the scaling factors are limited by other reasons and skewed.

fields also cause hot-carrier injection into the oxide leading to oxide charging and subsequent threshold-voltage shift and transconductance degradation.

These aforementioned phenomena will cause short-channel effects which can be summarized as follows: (1) V_T is not constant with L, (2) I_D does not saturate with V_D bias, both above and below threshold; (3) I_D is not proportional to $1/L$; and (4) device characteristics degrade with operation time. Because short-channel effects complicate device operation and degrade device performance, these effects should be eliminated or minimized so that a physical short-channel device can preserve the *electrical* long-channel behavior. In this section, we discuss MOSFET scaling and the short-channel effects that accompany device miniaturization. [Item-(3) is related to high-field mobility or velocity saturation, and has already been treated in Section 6.2.2.]

6.4.1 Device Scaling

The most-ideal scaling rule to avoid short-channel effects is simply to scale down all dimensions and voltages of a long-channel MOSFET so that the internal electric fields are kept the same.[41] This constant-field scaling is shown in Table 2 and Fig. 25.This approach offers a conceptually simple picture for device miniaturization. All dimensions, including channel length and width, oxide thickness, and junction

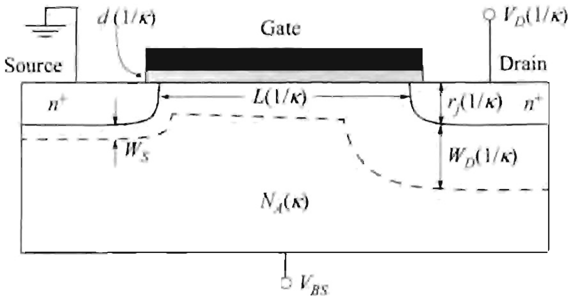


Fig. 25 Physical parameters for MOSFET scaling. Scaling factors for constant-field are indicated.

depth, are shrunk by the same *scaling factor* κ. The doping level is increased by κ, and all voltages are reduced by κ, leading to a reduction of the junction depletion width by about κ. Note that the subthreshold swing S remains essentially the same since S is proportional to $1 + C_D/C_{ox}$ and both capacitances are scaled up by the same factor κ.

Unfortunately such an ideal scaling rule is hindered by other factors that are fundamentally not scalable. First, the junction built-in voltage and the surface potential for the onset of weak inversion do not scale (only $\approx$ 10% change for 10 times increase in dopings). The range of gate voltage between depletion and strong inversion is approximately 0.5 V. These limitations stem from the fact that both the energy gap and thermal energy kT remain constant. The gate oxide thickness has the technological difficulty of defects as it approaches the low-nm scale. Tunneling through the oxide is another fundamental limitation. The quantum-mechanical effect discussed in Section 4.3.6 degrades the gate capacitance due to the fact that carriers are located at a finite distance ($\approx$1 nm) from the interface. The source and drain series resistance increases when r_j is decreased. This is especially detrimental when the current of the device increases at the same time. The channel doping cannot be increased indefinitely due to *p-n* junction breakdown. The threshold voltage cannot be scaled due to the off-current consideration, even with a fixed subthreshold swing. The supply voltage has been historically slow in scaling because of the system consideration, and also the push for higher speed. These limitations in scaling are summarized in Table 2, resulting in the actual nonideal scaling factors which are shown relative to the constant-field scaling. With these limitations, the field is no longer kept the same, and it increases with smaller gate lengths.

With the above practical limitations, other scaling rules have been proposed. These include constant-voltage scaling,[42] quasi-constant-voltage scaling,[43] and generalized scaling.[44] One other scaling rule with a unique feature of having flexible scaling factors has been proposed.[45] This allows the various device parameters to be adjusted independently as long as the overall behavior is preserved. Therefore, all device parameters do not have to be scaled by the same factor κ. The expression for minimum channel length for which long-channel behavior can be observed is found to follow a simple empirical relation:[45]

$$L \geq C_1 [r_j d (W_S + W_D)^2]^{1/3} \tag{92}$$

where C_1 is a constant, and $W_S + W_D$ is the sum of the source and drain depletion widths in a one-dimensional abrupt junction formulation:

$$W_D = \sqrt{\frac{2\varepsilon_s}{qN_A}(V_D + \psi_{bi} - V_{BS})}\,. \tag{93}$$

For $V_D = 0$, W_D equals W_S. A variation of this rule is also presented in Ref. 46.

We have discussed the nonideal factors that hinder constant-field scaling, resulting in some form of a penalty. On the positive notes, there are a couple of disruptive technologies on the horizon that will help scaling. First, MOSFETs built on a three-dimensional structure with an ultra-thin body will effectively eliminate most the conduction path for punch-through, and the requirement on channel doping can be relaxed (see Section 6.5.5). Second, research activities looking for gate dielectrics with high dielectric constants have been intense. Such a high-K gate dielectric can relax the physical thickness, improving the defect density and reducing the field for tunnelling. Both of these technologies can help to circumvent or delay the short-channel effects for a particular generation of channel length.

6.4.2 Charge Sharing from Source/Drain

Analysis of the channel charge so far is 1-dimensional, that is, both the inversion charge and depletion charge is completely balanced by the charge on the gate, so they can be treated as charge density. Detailed 2-dimensional examination at the ends of the channel reveals that some of the depletion charge is balanced by the n^+-source and drain, as shown in Fig. 26a. Departure from long-channel behavior can be shown to happen by applying the charge conservation principle to the region bounded by the gate, the channel, and the source/drain,[47]

$$Q'_M + Q'_n + Q'_B = 0, \tag{94}$$

where Q'_M is the total charge on the gate, Q'_n is the total inversion-layer charge, and Q'_B is the total ionized impurity in the depletion region. This of course assumes that all oxide and interface charges are zero. The threshold voltage, which can be viewed as voltage required to deplete the total bulk charge Q'_B in the maximum depletion width, is given by

$$V_T = V_{FB} + 2\psi_B + \frac{Q'_B}{C_{ox}A}, \tag{95}$$

where A is the gate area $Z \times L$. For long-channel devices, $Q'_B = qAN_AW_{Dm}$, where W_{Dm} is the maximum depletion-layer width,

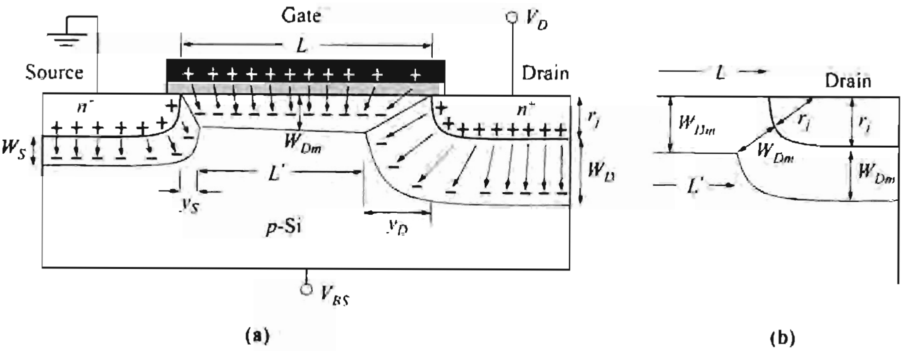


Fig. 26 Charge-conservation model for (a) $V_D > 0$, and (b) $V_D = 0$ where $W_D \approx W_S \approx W_{Dm}$. (After Ref. 47.)

$$W_{Dm} = \sqrt{\frac{2\varepsilon_s(2\psi_B - V_{BS})}{qN_A}}, \tag{96}$$

and 1-D analysis is sufficient.

For short-channel devices, the full effect of Q'_B on the threshold voltage is reduced, because near the source and drain ends of the channel, some field lines originating from the bulk charges under the channel region terminate at the source or drain instead of the gate (Fig. 26a). Note that the horizontal depletion-layer widths y_S and y_D are smaller than the vertical depletion-layer widths W_S and W_D, respectively, because the transverse field strongly influences the potential distribution at the surface.

First-order estimation of the threshold voltage can be made by considering the charge partition. The total bulk depletion charge can be estimated by the trapezoid [47]

$$Q'_B = ZqN_A W_{Dm}\left(\frac{L+L'}{2}\right). \tag{97}$$

For small drain bias, we can assume that $W_D \approx W_S \approx W_{Dm}$, and by straightforward trigonometric analysis (Fig. 26b),

$$L' = L - 2\left(\sqrt{r_j^2 + 2W_{Dm}r_j} - r_j\right). \tag{98}$$

The threshold-voltage shift from long-channel behavior is then given by

$$\begin{aligned}\Delta V_T &= \frac{1}{C_{ox}}\left(\frac{Q'_B}{ZL} - qN_A W_{Dm}\right) = -\frac{qN_A W_{Dm}}{C_{ox}}\left(1 - \frac{L+L'}{2L}\right)\\ &= -\frac{qN_A W_{Dm} r_j}{C_{ox}L}\left(\sqrt{1+\frac{2W_{Dm}}{r_j}} - 1\right).\end{aligned} \tag{99}$$

The negative sign means V_T is lowered and the transistor is easier to turn on. To take into account the effect of the drain voltage and the substrate bias, Eq. 99 can be modified to read [48]

$$\Delta V_T = -\frac{qN_A W_{Dm} r_j}{2C_{ox}L}\left[\left(\sqrt{1+\frac{2y_S}{r_j}} - 1\right) + \left(\sqrt{1+\frac{2y_D}{r_j}} - 1\right)\right], \tag{100}$$

where y_S and y_D are given as

$$y_S \approx \sqrt{\frac{2\varepsilon_s}{qN_A}(\psi_{bi} - \psi_s - V_{BS})}, \tag{101a}$$

$$y_D \approx \sqrt{\frac{2\varepsilon_s}{qN_A}(\psi_{bi} + V_D - \psi_s - V_{BS})}. \tag{101b}$$

Note that the threshold voltage becomes a function of both L and V_D. Figure 27 shows this dependence on both channel length and drain bias.

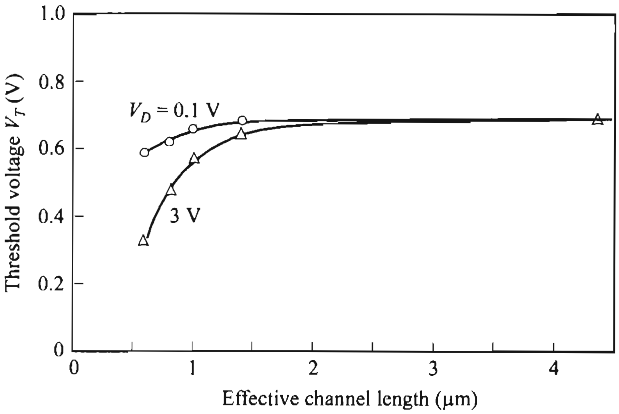


Fig. 27 Dependence of threshold voltage on channel length and drain bias. (After Ref. 49.)

6.4.3 Channel-Length Modulation

Figure 26a also shows that y_D is a high-field region where carriers are swept out efficiently. y_S is a transitional region where the electron concentration is higher than that in the main channel. So for consideration of the channel drift region, the *effective* channel length is more meaningful, given by

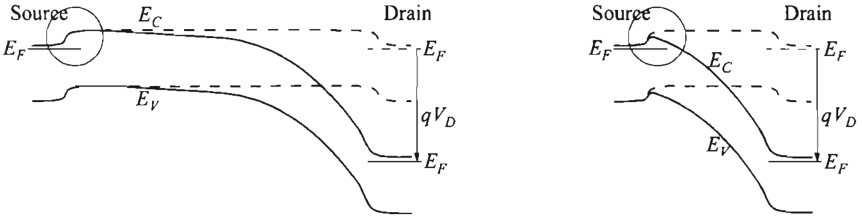
$$L_{\text{eff}} = L' = L - y_S - y_D. \tag{102}$$

This factor contributes to a drain-bias dependence of the effective channel length and partially accounts for the nonsaturating current with drain bias. Nevertheless, the change of channel length affects the current only linearly, whereas the barrier lowering caused by the drain bias, considered next, is much more pronounced since the current has an exponential dependence on the barrier.

6.4.4 Drain-Induced Barrier Lowering (DIBL)

We have pointed out that when the source and drain depletion regions are a substantial fraction of the channel length, short-channel effects start to occur. In extreme cases when the sum of these depletion widths approaches the channel length ($y_S + y_D \approx L$), more-serious effects will happen. This condition is commonly called punch-through. The net result is a large leakage current between the source and drain, and that this current is a strong function of the drain bias.

The origin of punch-through is the lowering of the barrier near the source, commonly referred to as DIBL (drain-induced barrier lowering). When the drain is close to the source, the drain bias can influence the barrier at the source end, such that the channel carrier concentration at that location is no longer fixed. This is demonstrated by the energy bands along the semiconductor surface, shown in Fig. 28. For a long-channel device, a drain bias can change the effective channel length, but the barrier at



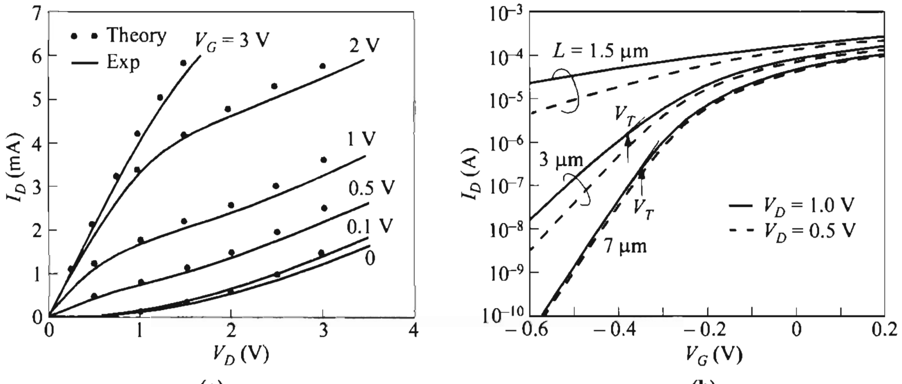
(a) (b)

Fig. 28 Energy-band diagram at the semiconductor surface from source to drain, for (a) long-channel and (b) short-channel MOSFETs, showing the DIBL effect in the latter. Dashed lines $V_D = 0$. Solid lines $V_D > 0$.

the source end remains constant. For a short-channel device, this same barrier is no longer fixed. The lowering of the source barrier causes an injection of extra carriers, thereby increasing the current substantially. This increase of current shows up in both above-threshold and subthreshold regimes.

Figure 28 shows that punch-through condition occurs at the semiconductor surface. In practical devices, it is common that the substrate concentration is reduced below the depth of the source/drain junction r_j. A reduced substrate doping widens the depletion widths so punch-through can also happen via a path in the bulk.

An example of severe punch-through characteristics above threshold is shown in Fig. 29a. For this device, at $V_D = 0$ the sum of y_S and y_D is 0.26 μm which is larger than the channel length of 0.23 μm. Therefore, the depletion region of the drain junc-



(a) (b)

Fig. 29 Drain characteristics of MOSFETs showing DIBL effect. (a) Above threshold. $L = 0.23$ μm. $d = 25.8$ nm. $N_A = 7\times10^{16}$ cm^{-3}. (b) Below threshold. $d = 13$ nm. $N_A = 10^{14}$ cm^{-3}. (After Ref. 50.)

tion has reached the depletion region of the source junction. Over the drain bias range shown, the device is operated in punch-through condition. Under such a condition, majority carriers in the source region (electrons in this case) can be injected into the depleted channel region, where they will be swept by the field and collected at the drain. The punch-through drain voltage can be estimated by the depletion approximation to be

$$V_{pt} \approx \frac{qN_A(L-y_S)^2}{2\varepsilon_s} - \psi_{bi}. \tag{103}$$

Drain current will be dominated by the space-charge-limited current:

$$I_D \approx \frac{9\varepsilon_s \mu_n A V_D^2}{8L^3} \tag{104}$$

where A is the cross-sectional area of the punch-through path. The space-charge-limited current increases with V_D^2 and is parallel to the inversion-layer current. The calculated points in the figure are from a 2-dimensional computer calculation incorporating the punch-through effect and field-dependent mobility effect.

The DIBL effect on subthreshold current is shown in Fig. 29b, for various channel lengths. The device with a 7-μm channel length shows long-channel behavior, that is, the subthreshold drain current is independent of drain voltage when $V_D > 3kT/q$ (Eq. 57). For $L = 3$ μm, there is a substantial dependence of current on V_D, with a corresponding shift of V_T (which is at the point of current departure of the I-V characteristic from the straight line). The subthreshold swing also increases. For an even shorter channel, $L = 1.5$ μm, long-channel behavior is totally lost. The subthreshold swing becomes much worst and the device cannot be turned off any more.

6.4.5 Multiplication and Oxide Reliability

We pointed out earlier that due to nonideal scaling, the internal electric fields in MOSFETs would increase with shorter channel lengths. In this section we discuss the anomalous currents associated with high fields, as well as their impacts. Figure 30 depicts all parasitic currents in addition to the main channel current. Note that the highest field occurs near the drain, and this is the location where most of the anomalous currents originate.

First, as the channel carriers (electrons) go through the high-field region, they acquire extra energy from the field without losing it to the lattice. These energetic carriers are called *hot carriers* whose kinetic energy is measured above the conduction band E_C. This extra energy, if larger than the Si/SiO$_2$ barrier (3.1 eV), enables them to escape to the oxide layer and to the gate terminal, and gives rise to a gate current.

Another major phenomenon happening in the high-field region is impact ionization which generates extra electron-hole pairs. These extra electrons go directly to the drain and add to the channel current. The path of the generated holes, however, is more diverse. A small fraction of them are driven to the gate, analogous to the hot electrons mentioned before. The majority of the generated holes flow to the substrate. For short-channel devices, some holes will go to the source. The division of these

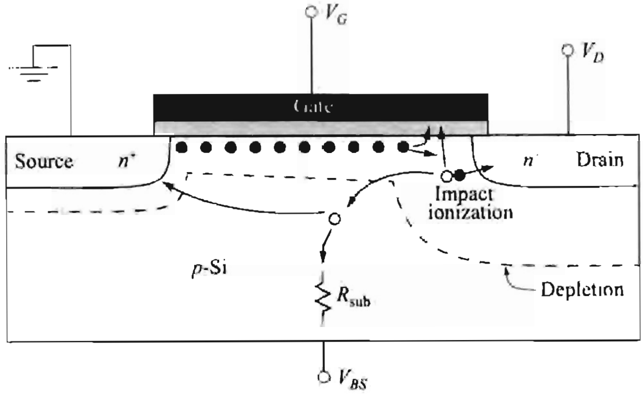


Fig. 30 Current components of a MOSFET under high fields.

paths depends on how good the substrate tie is. A perfect substrate tie ($R_{sub} = 0$) will sink all the hot holes and none of them will go to the source. As explained later, holes going to the gate or source will produce undesirable effects.

An example of the MOSFET terminal currents, including the gate current and substrate current, are shown in Fig. 31. Note that the gate current is due to hot electrons and hot holes *over* the barrier and is different from carriers tunneling *through* the barrier. This hot-carrier gate current peaks approximately at $V_G \approx V_D$. It is gener-

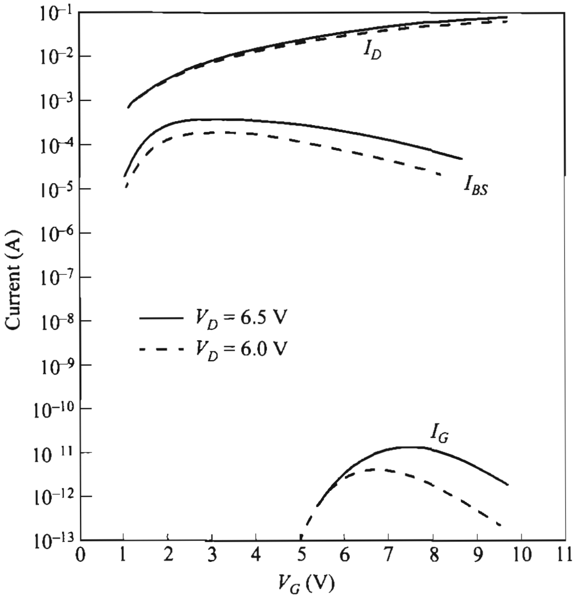


Fig. 31 Drain current, substrate current, and gate current vs. gate voltage of a MOSFET. $L/W = 0.8/30$ μm. (After Ref. 51.)

ally very small, and it by itself does not post a problem since it is negligible compared to the channel current. The impacts are the damages it creates. It is well known that these hot carriers create oxide charges and interface traps.[52] As a result, device characteristics change with operation time. In particular, the threshold voltage shifts, generally to a higher value, and the transconductance g_m degrades because of interface traps and reduced channel mobility. The subthreshold swing becomes larger because of increased interface-trap density. To reduce oxide charging, the density of water-related traps in the oxide should be minimized,[53] because such traps are known to capture hot carriers on their passage. In order to ensure MOSFETs' performance over a reasonable time, it is important to quantify the device lifetime over which the device parameters do not drift outside a given range, for a given bias condition. This is part of the specification of a MOSFET technology.

Figure 31 also shows the general characteristics of the substrate current. The substrate current has a unique bell shape as a function of gate bias.[54] It increases first with V_G, reaches a maximum then decreases. This maximum occurs usually around $V_G \approx V_D/2$. The maximum in I_{BS} can be explained as follows. Assuming that the impact ionization occurs uniformly in the high-field region, the substrate current can be written as

$$I_{BS} \approx I_D \alpha(\mathscr{E}) y_D, \tag{105}$$

where α is the ionization coefficient, the number of electron-hole pairs generated per unit distance, and is a strong function of the electric field; and y_D is the high-field or pinch-off region. For a given V_D, as V_G increases, both I_D and $V_{D\text{sat}}$ increase ($V_{D\text{sat}} \approx V_G - V_T$). When $V_{D\text{sat}}$ increases, the lateral field [$\approx (V_D - V_{D\text{sat}})/y_D$] decreases, causing a reduction of α. Thus we have two conflicting factors. The initial increase of I_{BS} is caused by the increase of drain current with V_G, and at larger V_G, the decrease of I_{BS} is due to the decrease of α. Maximum I_{BS} occurs where the two factors balance each other. Empirically, the substrate current can be expressed as

$$I_{BS} = C_2 I_D (V_D - V_{D\text{sat}}) \exp\left(\frac{-C_3}{V_D - V_{D\text{sat}}}\right), \tag{106}$$

where C_2 and C_3 are constants.

For a short-channel device, that is, a small source-drain separation, avalanched holes have increasing tendency to flow to the source.[55] This hole current constitutes a base current of a parasitic *n-p-n* bipolar-transistor action, where source-substrate-drain is equivalent to the emitter-base-collector configuration. For every hole that goes to the source, there are many electrons injected to the substrate. These electrons will be collected by the drain and show up as additional drain current. This bipolar current gain I_n/I_p is roughly determined by the ratio of N_D/N_A. Another way to look at this is that the substrate current develops a substrate voltage $I_{BS} \times R_{\text{sub}}$ which puts the source-substrate *p-n* junction under forward bias, thereby injecting electrons into the substrate. An additional effect is that the higher substrate potential lowers the threshold voltage for the surface channel and increases the surface-channel current. Both effects will increase the total drain current. These effects will become worst with an increase of R_{sub} and shorter L. In the extreme case of a MOSFET without a substrate contact ($R_{\text{sub}} = \infty$) such as an SOI or TFT structure (Section 6.5.4), the output curves show sudden rise of I_D with V_D. This is referred to as *kink effect* in the output characteristics.

In more-severe cases, the substrate current can induce source-drain breakdown from this parasitic *n-p-n* bipolar action. Analogous to the analysis of a bipolar breakdown (Section 5.2.3), since the source-drain distance is much shorter than the drain-to substrate contact, the base can be treated as open, and the MOSFET source-drain breakdown is given by the parasitic open-base bipolar breakdown (Eq. 47 of Chapter 5)

$$V_{BDS} = V_{BDx}(1 - \alpha_{npn})^{1/n}. \tag{107}$$

Here V_{BDx} is the drain-substrate *p-n* junction breakdown voltage, and n describes the shape of this diode breakdown characteristics. α_{npn} is the common-base current gain, given by the product of the base transport factor α_T and emitter efficiency γ. Assuming $\gamma \approx 1$, we have

$$\begin{aligned}\alpha_{npn} &= \alpha_T\gamma \approx \alpha_T \\ &\approx 1 - \frac{L^2}{2L_n^2}\end{aligned} \tag{108}$$

where L (channel length) is the effective base width, and L_n is the electron diffusion length in the substrate. From the above equations the breakdown voltage between source and drain is obtained for a short-channel MOSFET as

$$V_{BDS} \approx \frac{V_{BDx}}{2^{1/n}}\left(\frac{L}{L_n}\right)^{2/n}. \tag{109}$$

Equation 109 has been used to fit the data quite well, provided that the factor n is chosen to be 5.4.[55] The difference in breakdown voltage for different junction curvature can be explained by the dependence of V_{BDx} on r_j, as discussed in Chapter 2. To reduce the parasitic transistor effect, the resistance of the substrate R_{sub} should be minimized so that the product of $I_{BS} \times R_{\text{sub}}$ remains smaller than ≈ 0.6 V. Then the breakdown voltage of a short-channel MOSFET will no longer be limited by this parasitic bipolar effect, and higher voltages and more-reliable operation can be expected.

The drain-gate overlap region forms a gated-diode structure. For a thin oxide together with an abrupt junction, avalanche can occur during certain bias condition and it results in a drain leakage current going to the substrate. Such gated-diode avalanche current is called gate-induced drain leakage (GIDL) and the mechanism has been discussed in more details in Section 2.4.3. For an n-channel device, with a fixed drain bias, the normal channel current decreases with decreasing gate bias into the subthreshold regime. At some gate bias, the drain current becomes the GIDL current, and it rises again with more negative gate bias. Very often, in short-channel devices, this GIDL current already exists at $V_G = 0$, imposing a leakage current component in their off-state.[56]

6.5 MOSFET STRUCTURES

Up to now, Si MOSFET has been the workhorse of the electronics industry. As such, the MOSFET channel length and other dimensions have been pushed to shrink for the benefits of performance and density (see Fig. 1). While there is much discussion on what dimensions are the scaling limits,[57] it is certainly true that device scaling is getting increasingly difficult and has diminishing return. There are many possible reasons for hitting the end of scaling. These include: sensitivity of statistical doping fluctuations and surface charges, various forms of short-channel effects, quantum confinement in inversion layer which places a limit on gate capacitance, source/drain series resistance, etc. Most-recent data suggest that channel length below 20 nm is feasible, even with planar technology.[58-59] However, for practical applications, the scaling limit is most likely around 10 nm, even for 3-dimensional structures.

Many device structures have been proposed to control short-channel effects and improve MOSFET performance. We shall now consider the MOSFET structure broken down here into separate parts: channel doping, gate stack, and source/drain design. This is followed by some representative device structures for ultimate performance and special purposes.

6.5.1 Channel Doping Profile

Figure 32 shows the schematic of a typical high-performance MOSFET structure based on planar technology. The channel doping profile has a peak level slightly below the semiconductor surface. Such a retrograde profile is achieved with ion implantation, often of multiple doses and energies. The low concentration at the surface has the advantages of higher mobility due to reduced normal field and low threshold voltage. The high peak concentration below the surface is to control punch-through and other short-channel effects. Lower concentration is typically below the junction depth. It reduces the junction capacitance as well as the substrate-bias effect on threshold voltage.

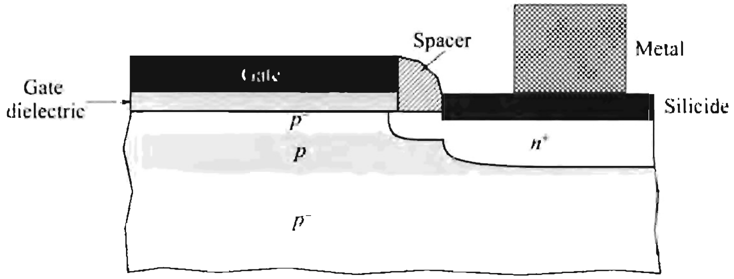


Fig. 32 High-performance MOSFET planar structure with a retrograde channel doping profile, two-step source/drain junction, and self-aligned silicide source/drain contact.

6.5.2 Gate Stack

The gate stack consists of the gate dielectric and the gate contact material. The gate dielectric has been exclusively SiO_2 right from the birth of MOSFET. In fact, the ideal Si-SiO_2 interface is the main factor responsible for the success of MOSFET. As the oxide thickness is scaled into the range below ≈ 2 nm, fundamental problem of tunneling and technological difficulty of defects start to demand alternatives. A sensible and popular solution that is actively sought after is a material with high dielectric constant, called high-K dielectric. Such a high-K dielectric can have a thicker physical thickness for the same capacitance, thus reducing its electric field and technological problem related to defects. With the value of dielectric constant considered, the common terminology used is the equivalent oxide thickness [EOT = thickness × $K(SiO_2)/K$]. Material options being examined are Al_2O_3, HfO_2, ZrO_2, Y_2O_3, La_2O_3, Ta_2O_5, and TiO_2. The dielectric constants for these materials range from 9 to 30, except for TiO_2 which is larger than 80. As seen [$K(SiO_2) = 3.9$], the EOT can easily be extended to below 1 nm if some of the options eventually are proven to be successful. Nevertheless, readers are reminded of the quantum-mechanical effect discussed in Section 4.3.6 which puts a limit on the gate capacitance.

The gate material has been polysilicon for a long time. The advantages of a poly-Si gate is its compatibility with the silicon processing, and its ability to withstand high-temperature anneal that is required after self-aligned source/drain implantation. Another important factor is that the work function can be varied by doping it into n-type and p-type. Such flexibility is crucial for a symmetric CMOS technology. One limitation of the poly-Si gate is its relatively high resistance. This does not result in penalty of dc characteristics since the gate terminates on the gate insulator. The penalty shows up in high-frequency parameters such as noise and f_{max} (Section 6.6.1). Another shortcoming of poly-Si gate is the finite depletion width at the oxide interface. This reduces the effective gate capacitance and becomes more severe with thinner oxides. To circumvent the problems of resistance and depletion, gate materials of silicides and metals are obvious choices. Potential candidates are TiN, TaN, W, Mo, and NiSi.

6.5.3 Source/Drain Design

Details of the source/drain structures are shown in Fig. 32. Typically the junction has two sections. The extension near the channel has shallower junction depth to minimize short-channel effects. Sometimes it is doped less heavily to reduce the lateral field for consideration of hot-carrier aging. For this purpose it is called a lightly doped drain (LDD). The deeper junction depth away from the channel helps to minimize the series resistance.

It has been pointed out that the sharpness or gradient of the source/drain profile is critical to minimize the series resistance.[60] We refer to Fig. 33 to understand its origin. In practice the profile is never perfectly abrupt, and there exists a region of accumulation layer (of n-type) before current spreads into the bulk of the

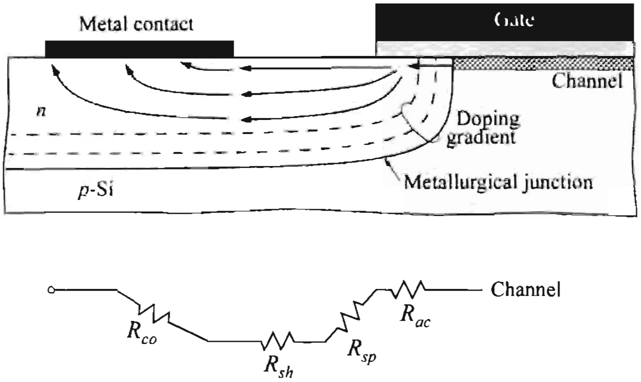


Fig. 33 Detailed analysis of different components of parasitic source/drain series resistance. R_{ac} is accumulation-layer resistance due to doping gradient. R_{sp} = spreading resistance. R_{sh} = sheet resistance. R_{co} = contact resistance. (After Ref. 60.)

source/drain. This accumulation-layer resistance R_{ac} is related to the transition distance before the doping reaches a critical level.

A major milestone for source/drain design is the development of the silicide contact technology which started in the early 1990s. Unlike the metal contact, the silicide can be made self-aligned to the gate, as shown in Fig. 32, thus minimizing the sheet-resistance component (R_{sh}) between the contact and the channel. In this way the silicide has become the metal contact because contact resistance between metal and silicide is very small. This *self-aligned silicide* process has been coined *salicide*. The salicide process is described is follows. After the gate definition, an insulator spacer is formed on the sides of the gate. A metal layer for silicidation is deposited uniformly, which at this stage is shorting the gate and the source/drain. After a thermal reaction at low temperature ($\approx 450°C$), the metal reacts with silicon to form silicide on the source/drain region. Silicide formation on the gate is optional depending on whether the gate is capped with an insulation layer as part of the gate stack. Metal over the spacer region and the field region (between transistors, not shown) remains metal since there is no exposed silicon for reaction. The metal is then removed with a selective chemical that etches metal only without etching silicide, thereby removing the shorting paths. Note that the silicide/silicon interface in Fig. 32 is slightly recessed. This is due to the consumption of silicon during silicide formation. Examples for salicides are $CoSi_2$, $NiSi_2$, $TiSi_2$, and PtSi.

Schottky-Barrier Source/Drain. Instead of *p-n* junction, use of Schottky-barrier contacts for the source and drain of a MOSFET can result in some advantages in fabrication and performance. Figure 34a shows a schematic MOSFET structure with such Schottky source and drain.[61] For a Schottky contact, the junction depth can effectively be made zero to minimize the short-channel effects. *n-p-n* bipolar-transistor action is also absent for undesirable effects such as bipolar breakdown and

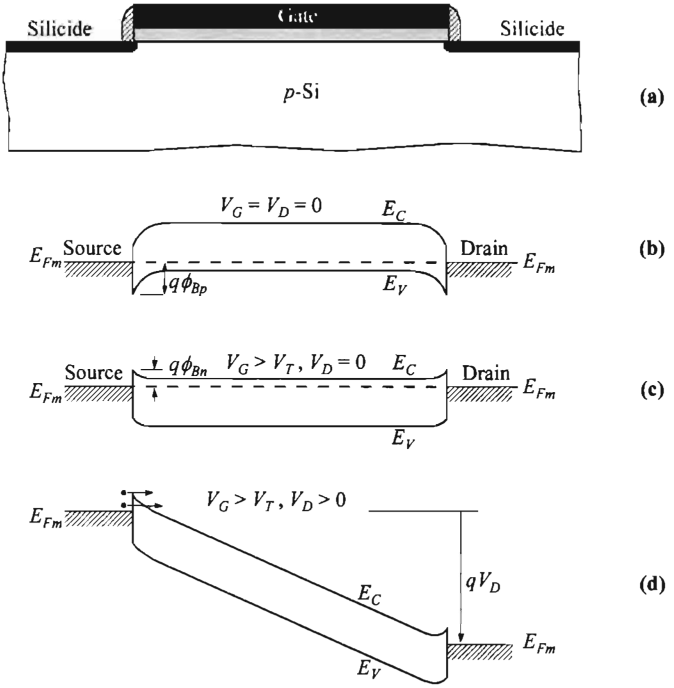


Fig. 34 MOSFET with Schottky-barrier source and drain. (a) Cross-sectional view of the device. (b)–(d) Band diagrams along semiconductor surface under various biases.

latch-up phenomenon[62] in CMOS circuits. Eliminating high-temperature implant anneal can promote better quality in the oxides and better control of geometry. In addition, this structure can be made on semiconductors such as CdS where *p-n* junctions cannot be easily formed.

Figures 34b–d show the working principle of a Schottky source/drain. At thermal equilibrium with $V_G = V_D = 0$, the barrier height of the metal to the *p*-substrate for holes is $q\phi_{Bp}$ (e.g., 0.84 eV for an ErSi-Si contact).[63] When the gate voltage is above threshold to invert the surface from *p*-type to *n*-type, the barrier height between the source and the inversion layer (electrons) is $q\phi_{Bn} = 0.28$ eV. Note that the source contact is reverse biased under operating conditions (Fig. 34d). For a 0.28-eV barrier, the thermionic-type reverse-saturation current density is of the order of 10^3 A/cm^2 at room temperature. To increase current density, metals should be chosen to give the highest majority-carrier barrier such that the minority-carrier barrier height is minimized. Additional current due to tunneling through the barrier should help to improve the supply of channel carriers. At the present, making the structure on a *p*-type Si substrate for *n*-channel MOSFET is more difficult compared to *p*-channel device with *n*-substrate, because metals and silicides that give large barrier heights on *p*-type silicon are less common.

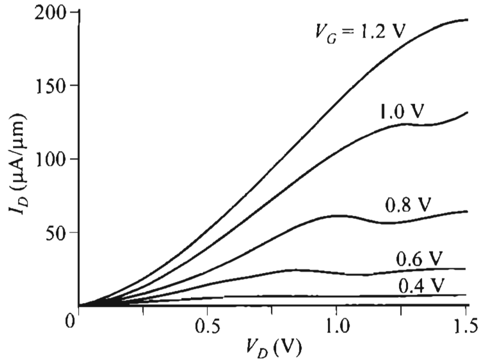


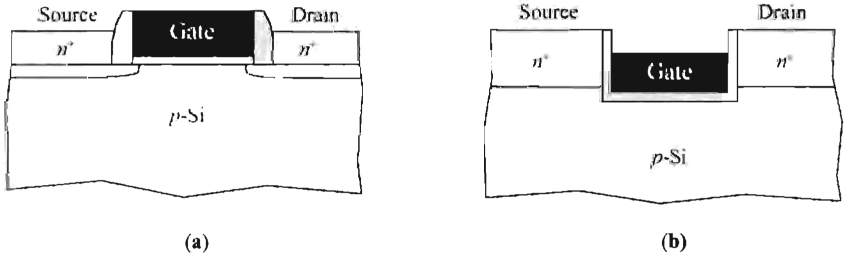
Fig. 35 *I-V* characteristics of *n*-channel MOSFET with Schottky source/drain. (After Ref. 63.)

The disadvantages of the Schottky source/drain are high series resistance due to the finite barrier height, and higher drain leakage current. Typical *I-V* curves show that current is starved at low-drain bias (Fig. 35). Also note that as shown in Fig. 34, the metal or silicide contact has to extend underneath the gate for continuity. This process is much more demanding than a junction source/drain which is done by self-aligned implantation and diffusion.

Raised Source/Drain. An advanced design is the raised source/drain where a heavily doped epitaxial layer is grown over the source/drain regions (Fig. 36a). The purpose is to minimize junction depth to control short-channel effects. Note that an extension underneath the spacer is still needed for continuity. An alternative is the recessed-channel MOSFET where the junction depth r_j is zero or negative (Fig. 36b).[64] The drawback of the recessed-channel structure, especially for submicron devices, is the difficulty in controlling the contour and the oxide thickness at the corners where the threshold voltage is determined. Also, oxide charging may be worsened because more hot-carrier injection will occur.

6.5.4 SOI and Thin-Film Transistor (TFT)

SOI. Unlike thin-film transistor, the top silicon layer of an SOI (silicon-on-insulator) wafer is high-quality single-crystalline material that is suitable for high-performance



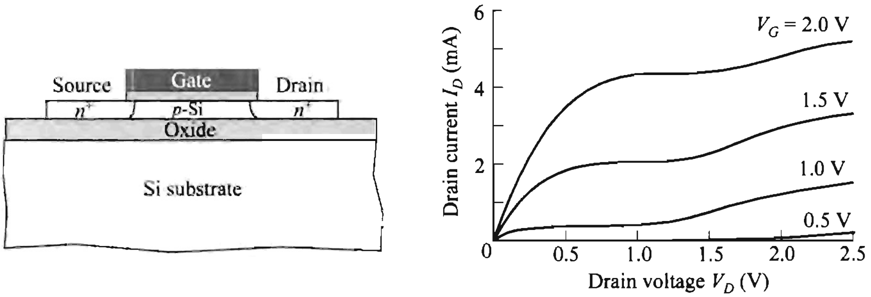
(a) (b)

Fig. 36 Means to reduce source/drain junction depth and series resistance. (a) Raised source/drain. (b) Recessed channel.

and high-density integrated circuits.[65] Many forms of SOI structures have been demonstrated with different insulator materials and holding substrates. These include silicon-on-oxide, silicon-on-sapphire (SOS), silicon-on-zirconia (SOZ), and silicon-on-nothing(air gap). In SOS and SOZ technologies, a single-crystalline silicon film is epitaxially grown on a crystalline insulating substrate. In these cases, the insulators are the substrates themselves, Al_2O_3 in SOS and ZrO_2 in SOZ. The difficulties in these techniques are the material quality when the film gets thinner. The first option, using oxide as an insulator and another Si wafer as the holding substrate, is by far the most popular. There are many ways to fabricate this structure. Among them SIMOX (separation by implantation of oxygen) where high-dose oxygen is implanted onto a silicon wafer followed by high-temperature anneal to form the buried SiO_2 layer. Another technique involves bonding of two silicon wafers one of which has an oxidized layer followed by thinning or removal of the majority of the top wafer until a thin silicon layer is left. One technique uses lateral epitaxial growth of silicon over an oxide layer, starting from a seed opening to the substrate. Another uses laser recrystallization, transforming amorphous silicon deposited onto the oxide layer into single-crystalline material, or into poly-crystalline form with large grain size.

Figure 37a shows a schematic diagram of an *n*-channel MOSFET made on an SOI substrate, with its typical *I-V* characteristics shown in Fig. 37b. The kinks associated with floating body without a substrate tie are noticeable.

The advantages of the SOI substrate include improved MOSFET scaling due to its thin body. A thin body can alleviate most problem with punch-through such that the channel can be doped lightly. The subthreshold swing is known to be improved. The buried oxide layer serves as good isolation to reduce capacitance to the substrate, giving rise to higher speed. As shown in Fig. 37a, device isolation is much easier, simply by removing the surrounding thin film. This can significantly improve the circuit density. This type of isolation, as opposed to junction isolation in planar technology, also eliminates latch-up phenomenon in CMOS circuits. The disadvantages of SOI is higher wafer cost, potentially inferior material properties, the kink effect, and worse heat conduction because of the oxide layer.



(a)

(b)

Fig. 37 (a) Typical structure of MOSFET on SOI wafer, and (b) its drain characteristics. (After Ref. 66.)

Thin-Film Transistor (TFT). The thin-film transistor usually refers to MOSFET as opposed to other kinds of transistors. The structure is similar to MOSFET built on SOI with the exception that the active film is a deposited thin film and that the substrate can be of any form.[67] Because the semiconductor layer is formed by deposition, the amorphous material has more defects and imperfections than in single-crystalline semiconductors, resulting in more complicated transport processes in the TFT. To improve device performance, reproducibility, and reliability, the bulk and interface-trap densities must be reduced to reasonable levels. In a TFT, the current is always very limited due to lower mobility, and leakage current is always higher due to defects. The main applications lie in the areas where a large-area or flexible substrate is required and conventional semiconductor processing is not feasible. A good example is large-area display where an array of transistors is required to control the array of lighting elements. In such applications, device performance such as current or speed is not critical.

6.5.5 Three-Dimensional Structures

In device scaling, the optimum design is with MOSFET built on a body of ultra-thin layer such that the body is fully depleted under the whole bias range. A design to achieve this more efficiently is to have a surround gate structure that encloses the body layer from at least two sides. Two examples of these 3-dimensional structures are shown in Fig. 38. They can be classified according to their current-flow pattern; the horizontal transistor[68–69] and vertical transistor.[70] While both of these are very challenging from a fabrication point of view, the horizontal scheme is more compatible with SOI technology and more data are reported in the literature. A new set of difficulties arise from the fact that the majority or all of the channel surface is on a vertical wall, for both of these structures. This fact presents great challenges in achieving a smooth channel surface from etching and growth or deposition of gate dielectrics on these surfaces. Formation of the source/drain junction is no longer

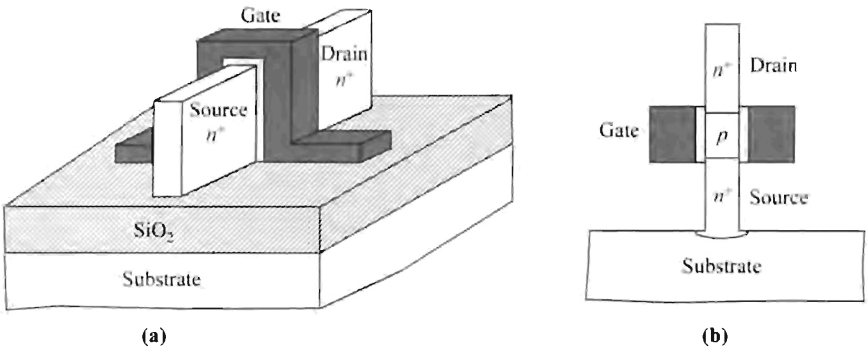


Fig. 38 Schematic 3-dimensional MOSFETs. (a) Horizontal structure. (b) Vertical structure. Note commonality of surround gate and thin body.

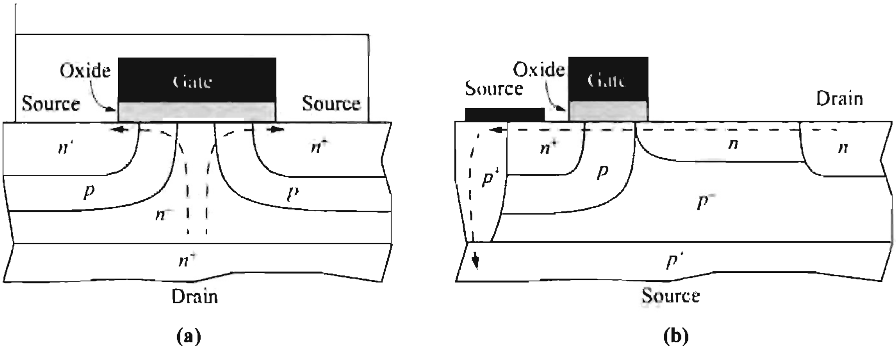
trivial by means of ion implantation. Salicide formation will also be much more difficult. Whether one of these turns out to be the device choice in the future remains to be seen.

6.5.6 Power MOSFETs

In general, power MOSFETs employ thicker oxides, deeper junctions, and have longer channel lengths. These generally post a penalty on device performance such as transconductance (g_m) and speed (f_T). Nevertheless, power applications from MOSFETs have been on the rise, for example, due to the increasing demand of cellular phones and cellular base stations which require extra-high voltage. We will present two power structures that are designed for RF power applications.

DMOS. As the name implies, in the DMOS (double-diffused MOS) transistor shown in Fig. 39a the channel length is determined by the higher diffusion rate of the p-dopant (e.g., boron) compared to the n^+-dopant (e.g., phosphorus) of the source. This technique can yield very short channels without depending on a lithographic mask. The p-diffusion serves as channel doping and has good punch-through control. The channel is followed by a lightly doped n^--drift region. This drift region is long compared to the channel, and it minimizes the peak electric field in this region by maintaining a uniform field.[71] Usually the drain is located at the substrate contact. The field near the drain is the same as in the drift region, so avalanche breakdown, multiplication, and oxide charging are lessened compared to conventional MOSFETs.

However, the threshold voltage V_T is more difficult to control in a DMOS transistor because the channel doping is no longer constant along its length.[72] Since V_T is determined by the local doping concentration along the semiconductor surface, varying doping level leads to variations in V_T as a function of distance and bias. Also the localization of punch-through control by a thin p-shield region requires a higher



(a) (b)

Fig. 39 (a) Vertical DMOS transistor and (b) LDMOS transistor. Current path is indicated by dashed line. In LDMOS transistors, it is common to connect the source to the substrate to reduce inductance of the bonding wire.

doping level compared to a conventional structure and it leads to poorer turn-off behavior for DMOS transistors.

LDMOS. The major difference of the LDMOS (laterally diffused MOS) transistor (Fig. 39b) from a DMOS transistor is that it has a lateral current-flow pattern. The drift region here is an implanted horizontal region. Such a horizontal arrangement enables the p^+-substrate to deplete this drift region at high drain bias. Yet at low drain bias its higher doping gives lower series resistance. This drift region, thus, behaves as a nonlinear resistor. At low drain bias, its resistance is determined by $1/nq\mu$. At high drain bias, this region is fully depleted so a large voltage drop can be supported. This concept is called RESURF (reduced surface field) technology.[73] Because of this feature, the drift region can be doped with a higher concentration compared to the DMOS transistor for a lower on-resistance. Another advantage of the LDMOS transistor is that the source can be tied internally to the substrate by a deep p-type diffusion. This avoids using a bond wire that has high inductance to the source. The LDMOS transistor can thus perform at higher speed.

6.6 CIRCUIT APPLICATIONS

6.6.1 Equivalent Circuit and Microwave Performance

The MOSFET is ideally a transconductance amplifier with an infinite input resistance and a current generator at the output. In practice, however, we have other nonideal circuit elements. An equivalent circuit is shown in Fig. 40 for the common-source connection. The gate resistance R_G is associated with the gate contact material over the oxide. The input resistance R_{in} is caused by tunneling current through the thin gate insulator, and it also includes any conductance through defects. This of course is a function of the oxide thickness. For a thermally grown silicon dioxide layer, this leakage current between the gate and the channel is negligibly small; thus, the input resistance is very high, one of the main advantages of a MOSFET. For oxides below

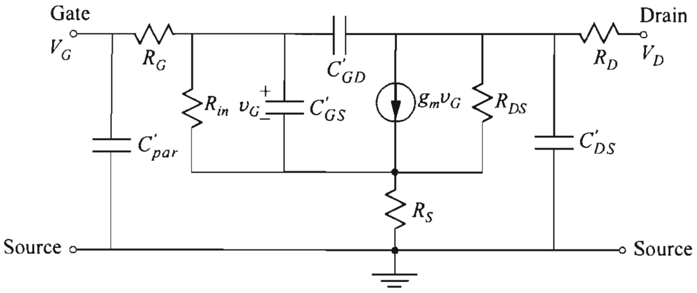


Fig. 40 Small-signal equivalent circuit of MOSFET for the common-source configuration. v_G is the small signal of V_G Capacitance symbols with prime designate total capacitance in Farad as opposed to capacitance per unit area.

a thickness of ≈ 5 nm, the tunneling current starts to become an important factor. The gate capacitance C'_G ($= C'_{GS} + C'_{GD}$) is mostly due to C_{ox} multiplied by the active channel area $Z \times L$. In practical devices, the gate extends somewhat above the source and drain regions, and these overlap capacitances add to the total C'_G. This fringing effect is also an important contribution to the feedback capacitance C'_{GD}. The drain output resistance R_{DS} is due to the fact that the drain current does not truly saturate with the drain bias. This effect is especially pronounced for short-channel devices, as part of the short-channel effects discussed earlier. The output capacitance C'_{DS} consists mostly of the two *p-n* junction capacitances connected in series through the semiconductor bulk.

In the saturation region, V_D and thus R_D has little effect on the drain saturation current. The R_S affects the effective gate bias, and the extrinsic transconductance is given by

$$g_{mx} = \frac{g_m}{1 + R_S g_m}. \tag{110}$$

In analyzing the microwave performance, we follow the same procedure as in Section 5.3.1 for obtaining the cutoff frequency f_T, defined as the frequency for unity current gain (ratio of drain current to gate current),

$$f_T = \frac{g_m}{2\pi(C'_G + C'_{par})} = \frac{g_m}{2\pi(ZLC_{ox} + C'_{par})}, \tag{111}$$

where C'_{par} is the total input parasitic capacitance. (See footnote* for a more-complete f_T expression with excessive R_S and R_D.) It is interesting to note that if C'_{par} is only due to the gate-drain and gate-source overlap regions, it has the same oxide-thickness dependence as g_m, and f_T would be independent of C_{ox} or oxide thickness. In the ideal case that there is zero parasitics, it can be shown that

$$\begin{aligned} f_T &= \frac{g_m}{2\pi ZLC_{ox}} \\ &= \frac{\upsilon}{2\pi L} = \frac{1}{2\pi\tau_t} \end{aligned} \tag{112}$$

where τ_t is the transit time across the channel length. Such an ideal case is in practice impossible to achieve, but this equation gives an estimate of the upper limit of f_T by using $\upsilon = \upsilon_s$. Again, in this limit, f_T is also independent of oxide thickness or g_m. However, g_m is important in practical devices with parasitics.

---

* In cases of very large source and drain resistances, the more-complete expression is

$$f_T = \frac{g_m}{2\pi\left[C'_G\left(1 + \frac{R_D + R_S}{R_{DS}}\right) + C'_{GD} g_m (R_D + R_S) + C'_{par}\right]}.$$

Another figure-of-merit for microwave performance is the maximum frequency of oscillation $f_{\max}$, the frequency at which the unilateral gain becomes unity. It is given by

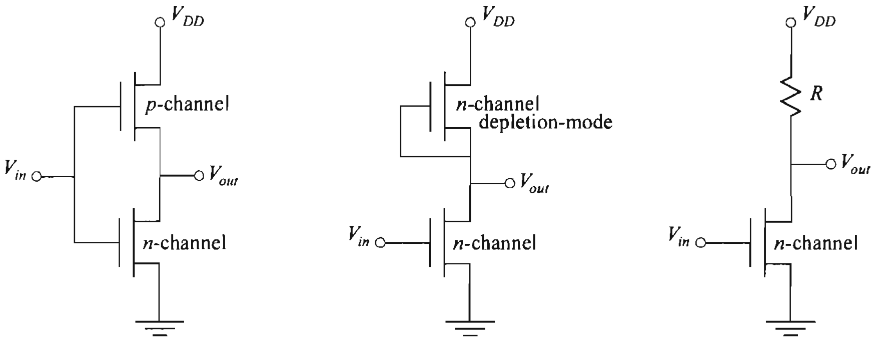
$$f_{\max} = \sqrt{\frac{f_T}{8\pi R_G C'_{GD}}}. \tag{113}$$

So for high-frequency performance, the most-important device parameters are g_m, R_G, and all other parasitic capacitances.

6.6.2 Basic Circuit Blocks

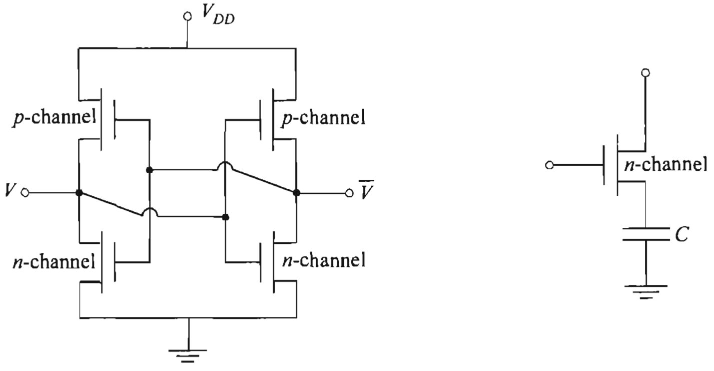
In this section we present the basic digital-circuit building blocks in both logic and memory circuits. The most-basic unit for a logic circuit is the inverter. Different configurations for MOSFET inverters are shown in Fig. 41. By far the most common is the CMOS (complementary MOS) inverter where both *n*-channel and *p*-channel transistors are used. This logic consumes very low dc power because when the input is either high or low, one of the transistors in series is off so that there is very little steady-state current (subthreshold current) passing through them. In fact, this is one of the main advantages and applications of MOSFETs where the insulated gate can withstand input voltage of any polarity. Such an arrangement is much more difficult with bipolar transistors or MESFETs without putting a large resistor in front of the input. In an NMOS logic (Fig. 41b), the load of the *p*-channel transistor is replaced with a depletion-mode *n*-channel transistor. The advantage is a simpler technology since a *p*-channel device is not required at the expense of higher dc power. This depletion-mode device with the gate tied to the source is basically a two-terminal nonlinear resistor, which is an improvement compared to a simple resistor load shown in Fig. 41c.

Two basic MOSFET memory cells, for SRAM (static random-access memory) and DRAM (dynamic random-assess memory) circuits, are shown in Fig. 42. The



(a) (b) (c)

Fig. 41 Versions of inverters: (a) CMOS logic. (b) NMOS logic with depletion-mode–transistor load, and (c) NMOS logic with resistor load.


V_{DD}
p-channel
p-channel
V
$\overline{V}$
n-channel
n-channel
n-channel
C

(a) (b)

Fig. 42 Basic memory cells in (a) SRAM and (b) DRAM circuits.

SRAM cell has two CMOS inverters connected back to back. It is a latch and a stable cell but it requires four transistors (six including controls for word line and bit line). The DRAM cell only requires one transistor and, thus, has very high memory density. Its memory information is stored as a charge across the capacitor. Since there is finite leakage of charge in the nonideal capacitor, the cell needs to be refreshed periodically, typically at a frequency of ≈ 100 Hz.

6.7 NONVOLATILE MEMORY DEVICES

Semiconductor memory devices are classified in Fig. 43. The first division is based on their ability to maintain their states when the power is disconnected. As the names imply, a volatile memory loses the content, but a nonvolatile memory does not need voltage to maintain the data.[74–76]

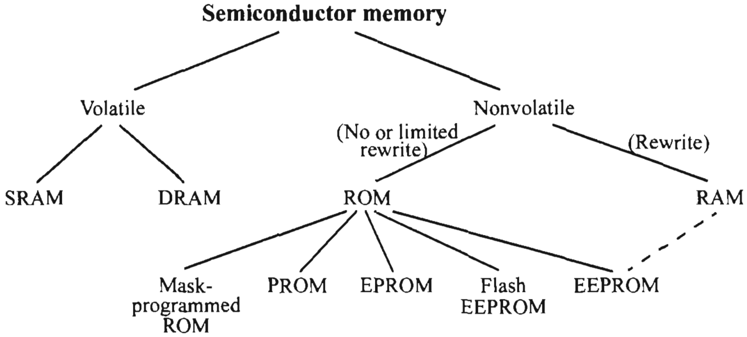

Semiconductor memory
Volatile
Nonvolatile
(No or limited rewrite)
(Rewrite)
SRAM
DRAM
ROM
RAM
Mask-programmed ROM
PROM
EPROM
Flash EEPROM
EEPROM

Fig. 43 Classification of semiconductor memories.

Before getting into each type of nonvolatile semiconductor memory, we should first clarify the difference between a *RAM* and a *ROM.* A RAM (random-access memory) has an *x-y* address for each cell, which distinguishes it from other serial memories such as magnetic memory. Strictly speaking, a ROM (read-only memory) also has random-access capability since the addressing architecture is similar. In fact the read processes of the RAM and ROM are almost identical. More appropriately, a RAM is sometimes called a read-write memory. However, the nonvolatile ROM has long started to develop some extent of rewriting capability. So the main difference now between a RAM and a ROM is the ease and frequency of erasing and programming. A RAM has almost equal opportunity of rewrite and read. A ROM in general has much more frequent read than rewrite. It itself has a spectrum of rewriting capability, ranging from a pure ROM without any writing capability to a full-feature EEPROM. Because a ROM is smaller in size and more cost-effective than a RAM, it is used whenever frequent rewriting is not required. With this background, different types of nonvolatile memories are explained below:

Mask-programmed ROM: The memory content is fixed by the manufacturer and is not programmable once it is fabricated. Sometimes mask-programmed ROM is simply referred to as ROM.

PROM: Programmable ROM is sometimes called field-programmable ROM or fusible-link ROM. The connectivity of the array is custom programmed using the technique of fusing or antifusing. After programming, the memory works as a ROM.

EPROM: In an electrically programmable ROM, programming is performed by hot-electron injection or tunneling to the floating gate, and it requires biases on both the drain and the control gate. Global erase is by exposure to a UV light or x-ray. Selective erase is not possible.

Flash EEPROM: A flash EEPROM, as opposed to a full-feature EEPROM below, can be erased electrically but only by a large block of cells simultaneously. It loses byte selectivity but maintains a one-transistor cell. It is, thus, a compromise between an EPROM and a full-feature EEPROM.

EEPROM: In an electrically erasable/programmable ROM, not only can it be erased electrically, but also selectively by byte address. To erase selectively, a select transistor is needed for each cell, leading to a two-transistor cell. This makes it less popular than a flash EEPROM.

Nonvolatile RAM: This memory can be viewed as a nonvolatile SRAM, or EEPROM with short programming time as well as high endurance. If technology allows the aforementioned features, this would be the ideal memory.

When the gate electrode of a conventional MOSFET is modified so that semipermanent charge storage inside the gate stack is possible, the new structure becomes a nonvolatile memory device. Since the first nonvolatile memory device proposed by Kahng and Sze in 1967,[77] various device structures have been made, and nonvolatile memory devices have been extensively used in commercial products. The two groups of nonvolatile memory devices are the floating-gate devices and the charge-trapping devices (Fig. 44). In both types of devices, charges are injected from the silicon substrate across the first insulator and stored in the floating gate or at the nitride-oxide

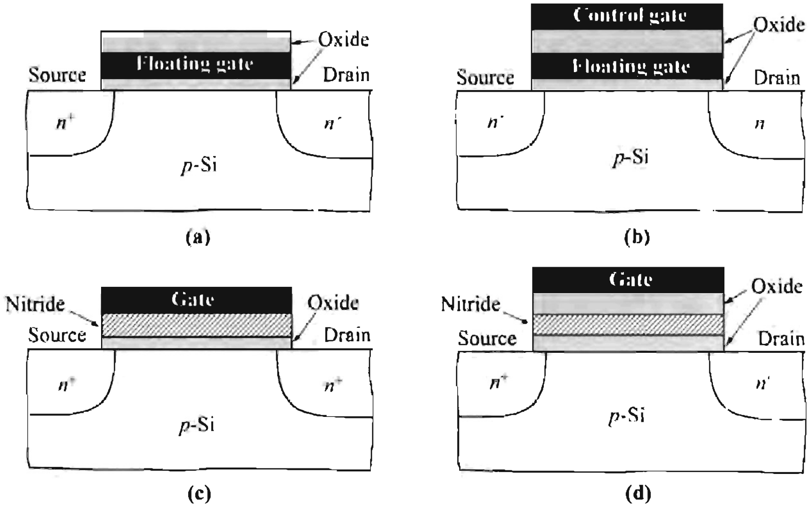


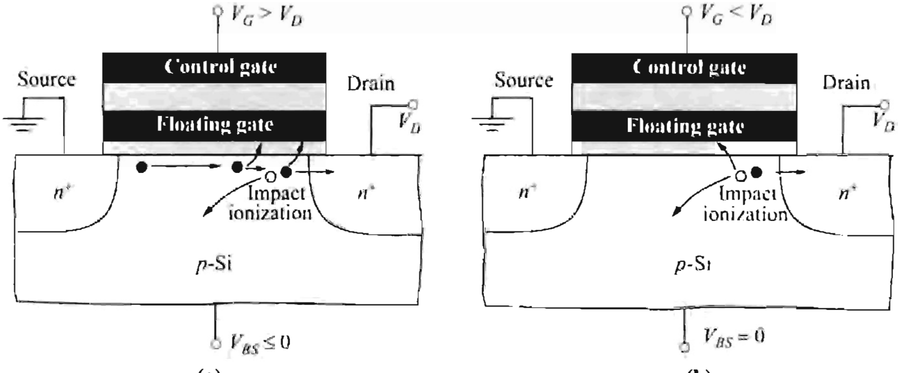
Fig. 44 Variations of nonvolatile memory devices: Floating-gate devices as (a) FAMOS transistor and (b) stacked-gate transistor. Charge-trapping devices as (c) MNOS transistor and (d) SONOS transistor.

interface. The stored charge gives rise to a threshold-voltage shift, and the device is at a *high-threshold state* (programmed). For a well-designed memory device, the charge retention time can be over 100 years. To return to the *low-threshold state* (erased), a gate voltage or other means (such as ultraviolet light) can be applied to erase the stored charge.

6.7.1 Floating-Gate Devices

In a floating-gate memory device, charge is injected to the floating gate to change the threshold voltage. The two modes of programming are hot-carrier injection and Fowler-Nordheim tunneling. Figure 45a shows the mechanisms of hot-carrier injection. Near the drain, the lateral field is at its highest level. The channel carriers (electrons) acquire energy from the field and become hot carriers. When their energy is higher than the barrier of the Si/SiO_2 interface, they can be injected to the floating gate. At the same time, the high field also induces impact ionization. These generated secondary hot electrons can also be injected to the floating gate. The hot-carrier injection currents give rise to the equivalence of gate current in a regular MOSFET, and is shown in Fig. 31. This gate current peaks at $V_{FG} \approx V_D$ where V_{FG} is the potential of the floating gate.

Figure 45b shows the original method of hot-carrier injection using drain-substrate avalanche. In this scheme, the floating-gate potential is more negative such that hot holes are injected instead.* This injection scheme is found to be less efficient and is no longer used in practice.



(a) (b)

Fig. 45 Charging of the floating gate by hot carriers. (a) Hot electrons from channel and impact ionization. (b) Hot holes from drain avalanche. Note difference in gate bias between the two figures.

Besides hot-carrier injection, electrons can be injected by tunneling. In this programming mode, the electric field across the bottom oxide layer is most critical. On application of a positive voltage V_G to the control gate, an electric field is established in each of the two insulators (Fig. 44b). We have, from Gauss' law, that

$$\varepsilon_1 \mathscr{E}_1 = \varepsilon_2 \mathscr{E}_2 + Q \tag{114}$$

and

$$V_G = V_1 + V_2 = d_1 \mathscr{E}_1 + d_2 \mathscr{E}_2 , \tag{115}$$

where the subscripts 1 and 2 correspond to the bottom and top oxide layer respectively, and Q (negative) is the stored charge on the floating gate. From Eqs. 114–115 we obtain

$$\mathscr{E}_1 = \frac{V_G}{d_1 + d_2(\varepsilon_1/\varepsilon_2)} + \frac{Q}{\varepsilon_1 + \varepsilon_2(d_1/d_2)} . \tag{116}$$

The current transport in insulators is generally a strong function of the electric field. When the transport is Fowler-Nordheim tunneling, the current density has the form

$$J = C_4 \mathscr{E}_1^2 \exp\left(\frac{-\mathscr{E}_0}{\mathscr{E}_1}\right) \tag{117}$$

where C_4 and $\mathscr{E}_0$ are constants in terms of effective mass and barrier height. This type of current transport occurs in SiO_2 and Al_2O_3 as discussed in Chapters 4 and 8.

---

* The original devices were p-channel such that hot electrons are injected under this scheme which are relatively more efficient than hot holes. We use the same n-channel for better comparison.

Using either hot-carrier injection or tunneling as programming mechanism, after charging, the total stored charge Q is equal to the integrated injection current since the gate is floating. This causes a shift of the threshold voltage by the amount

$$\Delta V_T = -\frac{d_2 Q}{\varepsilon_2}. \tag{118}$$

This threshold-voltage shift can be directly measured as shown in the I_D-V_G plot (Fig. 46). Alternately, the threshold-voltage shift can be measured from the drain conductance. The change in V_T results in a change in the channel conductance g_D of the MOSFET. For small drain voltages, the channel conductance of an n-channel MOSFET is given by

$$g_D = \frac{I_D}{V_D} = \frac{Z}{L}\mu C_{ox}(V_G - V_T), \qquad V_G > V_T. \tag{119}$$

After altering the charge on the floating gate by Q (negative charge), the g_D-V_G plot shifts to the right by ΔV_T.

To erase the stored charge, a negative bias is put on the control gate or a positive bias on the source/drain. The process is the reverse of the tunneling process described above, and the stored electrons tunnel out of the floating gate to the substrate.

The programming and erasing sequence of a floating-gate memory can be understood with the energy-band diagrams in Fig. 47. In Fig. 47b, electron injection can be due to hot carriers over the barrier or tunneling through the barrier. Figure 47c shows that the accumulated negative charge at the floating gate raises the threshold voltage compared to its initial condition in Fig. 47a. The erase is carried out by electron tunneling from the floating gate back to the substrate (Fig. 47d).

In both programming and erasing operations, it is important to modulate the floating-gate potential efficiently by the control-gate applied voltage. An important parameter in the floating-gate memory is the coupling ratio which determines the portion of the control-gate voltage that gets coupled to the floating gate capacitively. This coupling ratio is determined by the capacitance ratio

$$R_{CG} = \frac{C_2'}{C_1' + C_2'} \tag{120}$$

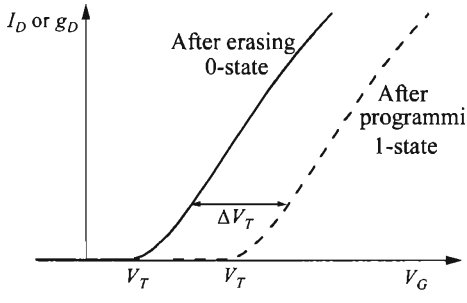


Fig. 46 Drain-current characteristics of a stacked-gate n-channel memory transistor, showing the change of threshold voltage after erasing and programming.

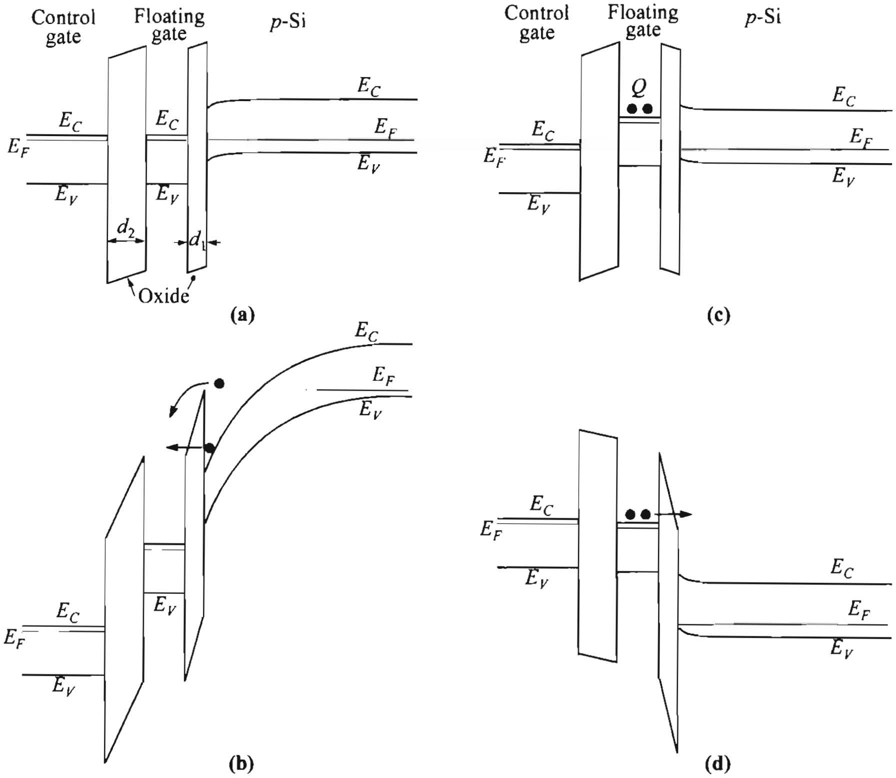

Control gate
Floating gate
p-Si
E_C
E_F
E_V
d_2
d_1
Oxide
(a)
Control gate
Floating gate
p-Si
Q
E_C
E_F
E_V
(c)
E_C
E_F
E_V
(b)
E_C
E_F
E_V
(d)

Fig. 47 Energy-band diagrams for a stacked-gate memory transistor at different stages of operation. (a) Initial stage. (b) Charging by hot electrons or electron tunneling. (c) After charging, the floating-gate having charge Q (negative) is at higher potential and V_T is increased. (d) Erasing by electron tunneling.

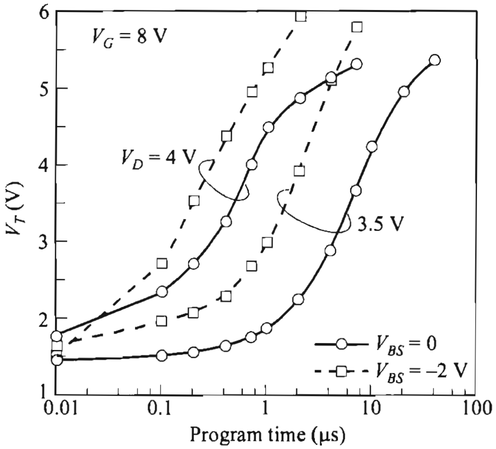


Fig. 48 Programming of floating-gate memory using hot-electron injection. (After Ref. 80.)

where C_1' and C_2' are the capacitances associated with the bottom and top insulator layers, respectively. Note that in practice, the areas of the control gate and the floating gate are not necessarily the same. More often than not, the top control gate wraps around the floating gate so the top capacitor has a larger area, unlike that shown in Figs. 44 and 45. The parameters C_1' and C_2' represent their total net capacitances. The floating-gate potential is given by

$$V_{FG} = R_{CG}V_G. \tag{121}$$

In practical devices, the bottom layer has a tunnel oxide of ≈ 80 Å, while the top insulator stack typically has an equivalent oxide thickness of ≈ 140 Å. A larger top area makes up for the difference in capacitance per unit area, and the coupling ratio is typically around 0.5–0.6.

In device structure, the first EPROM was developed using a heavily doped polysilicon as the floating-gate material (Fig. 44a). The device uses drain-substrate avalanche shown in Fig. 45b and is known as floating-gate avalanche-injection MOS (FAMOS) memory.[78] The polysilicon gate is embedded in oxide and is completely isolated. To inject charge into the floating gate (that is, to program), the drain junction is biased to avalanche breakdown, and holes in the avalanche plasma are injected from the drain region into the floating gate (*see footnote on p. 353). To erase the FAMOS memory, ultraviolet light or x-ray is used. Electrical erasing cannot be used because the device has no external gate.

To enable electrical erase, the stacked-gate structure with double-level polysilicon gates has been in popular use (Fig. 44b).[79] The external control gate makes electrical erasing possible and also improves the programming efficiency. An example of the programming transient based on hot-carrier injection is shown in Fig. 48.

In EEPROM circuits, it is more common to use tunneling as an injection mechanism for programming. A successful commercial device, called FLOTOX (floating-gate tunnel oxide) transistor, confines the tunneling process to a small area over the drain, as shown in Fig. 49. Typical programming and erasing transients for the FLOTOX transistors are shown n Fig. 50.

After programming, by definition, a long retention time is required for nonvolatile memory operation. The retention time is defined as the time when the stored charge decreases to 50% of its initial value and is expressed by

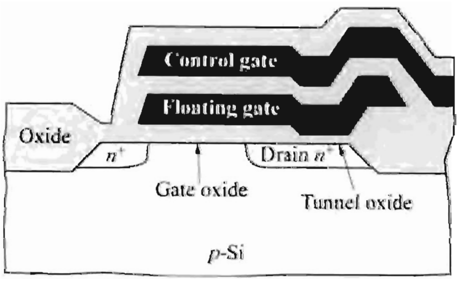


Fig. 49 Structure of the FLOTOX transistor which uses tunneling for both programming and erasing. (After Ref. 81.)

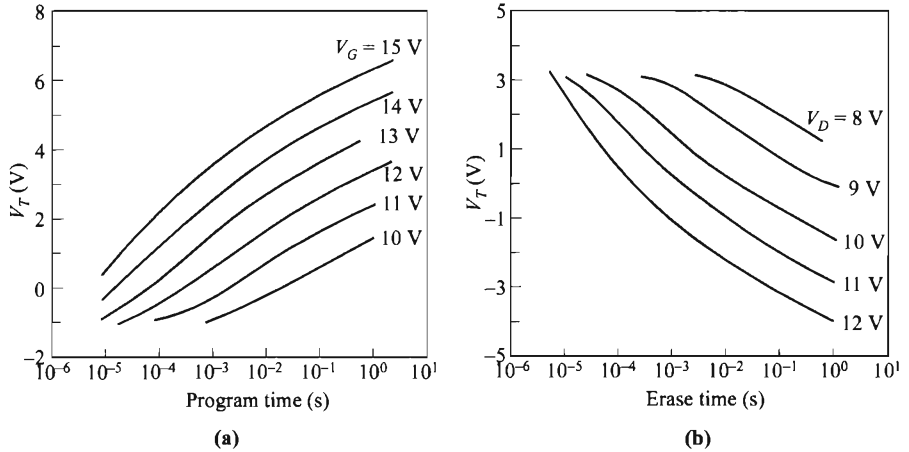


Fig. 50 Typical programming and erasing times for FLOTOX memory device. (After Ref. 76.)

$$t_R = \frac{\ln(2)}{\nu \exp(q\phi_B/kT)} \tag{122}$$

where ν is the dielectric relaxation frequency, and ϕ_B is the barrier height of the floating gate to oxide. The retention time is very sensitive to temperature. Typical retention times at 125°C and 170°C with $\phi_B = 1.7$ V are found to be about 100 years and 1 year respectively.[82]

6.7.2 Charge-Trapping Devices

MNOS Transistor. As a memory device, in the MNOS (metal-nitride-oxide-silicon) transistor, the silicon-nitride layer is used as an efficient material to trap electrons as current passes through the dielectric.[83] Other insulators in place of the silicon-nitride film such as aluminum oxide, tantalum oxide, and titanium oxide have been used but are not as common. Electrons are trapped in the nitride layer close to the oxide-nitride interface. The function of the oxide is to provide a good interface to the semiconductor and to prevent back-tunneling of the injected charge for better charge retention. Its thickness has to be balanced between retention time and programming voltage and time.

Figure 51 shows the basic band diagrams for the programming and erasing operations. In the programming process, a large positive bias is applied to the gate. Current conduction is known to be due to electrons that are emitted from the substrate to the gate. The conduction mechanisms in the two dielectric layers are very different and have to be considered in series. The current through the oxide J_{ox} is by tunneling. Notice that electrons tunnel through the trapezoidal oxide barrier, followed by a triangular barrier in the nitride. This form of tunneling has been identified as modified

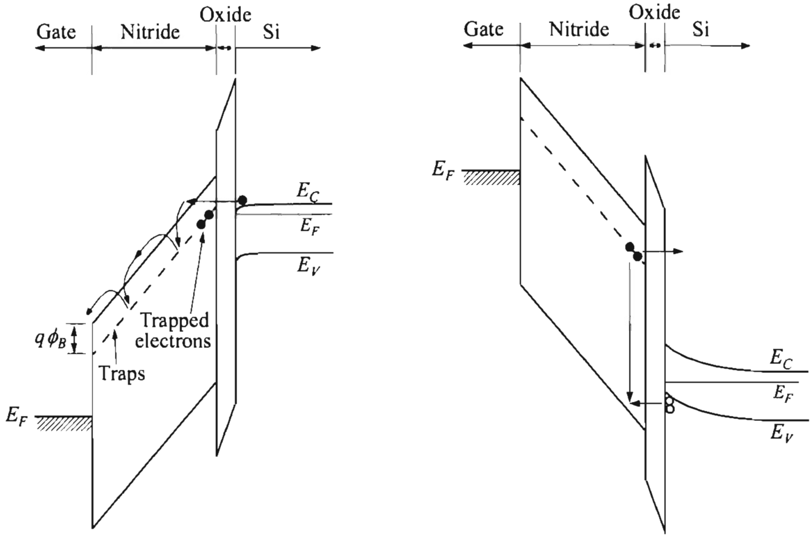


Fig. 51 Rewriting of MNOS memory. (a) Programming: electrons tunnel through oxide and are trapped in the nitride. (b) Erasing: holes tunnel through oxide to neutralize the trapped electrons, and tunneling of trapped electrons.

Fowler-Nordheim tunneling, as opposed to Fowler-Nordheim tunneling through a single triangular barrier. It has the following form of

$$J_{ox} = C_5 \mathscr{E}_{ox}^2 \exp\left(-\frac{C_6}{\mathscr{E}_{ox}}\right), \tag{123}$$

where $\mathscr{E}_{ox}$ is the field in the oxide layer, and C_5 and C_6 are constants. The current through the nitride layer J_n is controlled by Frenkel-Poole transport, which has the form

$$J_n = C_7 \mathscr{E}_n \exp\left[\frac{-q(\phi_B - \sqrt{q\mathscr{E}_n/\pi\varepsilon_n})}{kT}\right] \tag{124}$$

where $\mathscr{E}_n$ and ε_n are the electric field and permittivity in nitride, ϕ_B is the trap level below the conduction band ($\approx$ 1.3 V), and C_7 is a constant [= 3×10^{-9} $(\Omega\text{-cm})^{-1}$].

It is known that at the beginning of the programming process, modified Fowler-Nordheim tunneling is capable of a higher current, and current conduction is limited by Frenkel-Poole transport through the nitride layer. When the negative charge starts to build up, the oxide field decreases and the modified Fowler-Nordheim tunneling starts to limit the current. The threshold voltage as a function of programming pulse width is shown in Fig. 52. Initially, the threshold voltage changes linearly with time, followed by a logarithmic dependence, and finally it tends to saturate. This programming speed is largely affected by the choice of oxide thickness; a thinner oxide allows

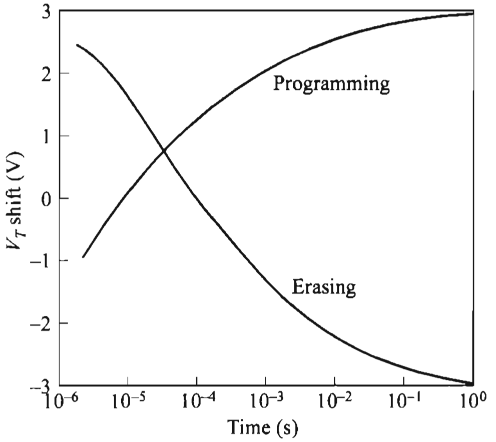


Fig. 52 Typical programming and erasing rates for MNOS transistor. (After Ref. 83.)

a shorter programming time. Programming speed has to be balanced with charge retention since too thin an oxide will allow the trapped charge to tunnel back to the silicon substrate.

The total gate capacitance C_G of the dual dielectrics is equal to the serial combination of their capacitances

$$C_G = \frac{1}{(1/C_n)+(1/C_{ox})} = \frac{C_{ox}C_n}{C_{ox}+C_n} , \tag{125}$$

where the capacitances $C_{ox} = \varepsilon_{ox}/d_{ox}$ and $C_n = \varepsilon_n/d_n$ correspond to the oxide and nitride layers respectively. The amount of trapped charge density Q near the nitride-oxide interface depends on the trapping efficiency of the nitride and is proportional to the integrated Frenkel-Poole current having passed through it. The final threshold-voltage shift is given by

$$\Delta V_T = -\frac{Q}{C_n} . \tag{126}$$

In the erasing process, a large negative bias is applied to the gate (Fig. 51b). Traditionally, the discharge process is believed due to the tunneling of trapped electrons back to the silicon substrate. New evidence shows that the major process is due to tunneling of holes from the substrate to neutralize the trapped electrons. The discharge process as a function of pulse width is also shown in Fig. 52.

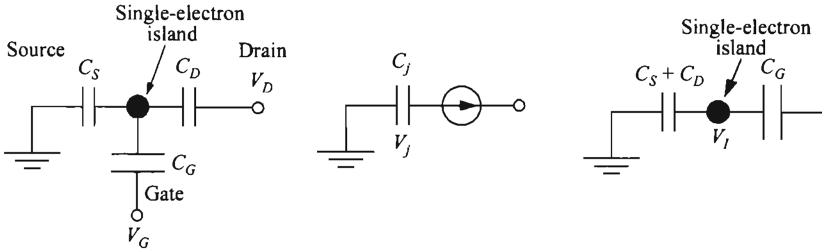
The advantages of the MNOS transistor include reasonable speed for programming and erasing, so it is a candidate as a nonvolatile RAM device. It also has superior radiation hardness, due to minimal oxide thickness and the absence of a floating gate. The drawbacks of the MNOS transistor are large programming and erasing voltages and nonuniform threshold voltage from device to device. The passage of tunneling current gradually increases the interface-trap density at the semiconductor surface and also causes a loss of trapping efficiency due to leakage or tunneling of trapped electrons back to the substrate. These result in a narrowing threshold voltage window after many cycles of programming and erasing. The major reliability problem of the MNOS transistor is the continuous loss of charge through the thin oxide. It should be pointed out that unlike a floating-gate structure, the programming current has to pass through the entire channel region, so that the trapped charge is distributed uniformly throughout the channel. In a floating-gate transistor, the charge injected to the floating gate can redistribute itself within the gate material, and injection can take place locally anywhere along the channel.

SONOS Transistor. The SONOS (silicon-oxide-nitride-oxide-silicon) transistor (Fig. 44d) is sometimes called the MONOS (metal-oxide-nitride-oxide-silicon) transistor. It is similar to an MNOS transistor except that it has an additional blocking oxide layer placed between the gate and the nitride layer, forming an ONO (oxide-nitride-oxide) stack. This top oxide layer is usually similar in thickness to the bottom oxide layer. The function of the blocking oxide is to prevent electron injection from the metal to the nitride layer during erase operation. As a result, a thinner nitride layer can be used, leading to lower programming voltage as well as better charge retention. The SONOS transistor now replaces the older MNOS configuration, but the operation principle remains the same.

6.8 SINGLE-ELECTRON TRANSISTOR

With the continuing advancement of technology to nano-scale device geometry, there are new experimental observations that have not been possible before. One of them is the charge-quantization effect in a single-electron transistor (SET),[84] first observed in 1987.[85] The structure of an SET is represented by the schematic circuit diagram in Fig. 53a. It has a central single-electron island that has to be extremely small. The island is connected between the source and drain via capacitors through which tunneling occurs to conduct current. The third terminal is the insulated gate and its purpose is to control the current between the source and drain, similar to the case of an FET.

The opportunity to observe quantization of charge comes directly from the small dimension of the single-electron island. The minimum energy needed to transport a



(a) (b) (c)

Fig. 53 Circuit representations of (a) single-electron transistor, (b) charging a tunneling capacitor, and (c) single-electron box.

single electron charge to and from the island is $q^2/2C_\Sigma$, where C_Σ is its total capacitance,

$$C_\Sigma = C_S + C_D + C_G. \tag{127}$$

This energy also must be much larger than the thermal energy for experimental observation, requiring that

$$\frac{q^2}{2C_\Sigma} > 100kT. \tag{128}$$

At room temperature, C_Σ needs to be on the order of aF (10^{-18} F). This necessitates a single-electron island size less than 1–2 nm. It is also interesting to note that this effect does not require the island to be made of semiconductor materials and most reported results are based on metal dots. Although for the limit of small islands such as quantum dots made out of semiconductors, since the total number of electrons within the dots (< 100) is much less than that in metal ($\approx 10^7$), a second effect of quantization of energy levels can also be observed. This causes addition structures in the I_D-V_D characteristics,[86] but does not contribute to the most-important features of an SET and will not be discussed here. As a matter of fact, the SET does not require any semiconductor material, but only needs metal and insulator.

The capacitors between the island and the source or drain are characterized by the tunneling resistances R_{TS} and R_{TD}. These resistances need to be small (thin insulator layers) to conduct a reasonable amount of current. But they are bound in the lower limit by the Uncertainty principle that electrons have to be treated as particles being clearly on either side of the junction. This requires that

$$R_{TS} \approx R_{TD} > \frac{h}{q^2} \tag{129}$$

(h/q^2 = 25.8 kΩ) and they should be above ≈ 1 MΩ.

The basic I-V characteristics of an SET are shown in Fig. 54. First, Fig. 54a shows that at most values of V_G, there is a knee V_D below which current is very much suppressed. This threshold drain voltage, caused by Coulomb blockade, is explained later. Another important feature is that this Coulomb blockade can be varied by the gate voltage. At some values of V_G, the Coulomb blockade totally vanishes. Shown in Fig. 54b, the cycle can be repeated many times and is called a Coulomb-blockade oscillation. This is very different from the gate control of a regular transistor, where the current can be turned on or off monotonically.

To explain these characteristics, it is best to go back to the simplest structure: a tunneling capacitor as shown in Fig. 53b. Here the capacitor is charged by a small current source, so the junction voltage V_j will increase until an electron can tunnel. The basis for the Coulomb blockade is that it requires a certain minimum V_j before there is enough energy for an electron to tunnel. The minimum energy needed is $q^2/2C_j$, which will be the change of energy of the capacitor when an electron tunnels. This is also the same as the energy gained by the electron when tunneling across the capacitor of voltage V_j, giving

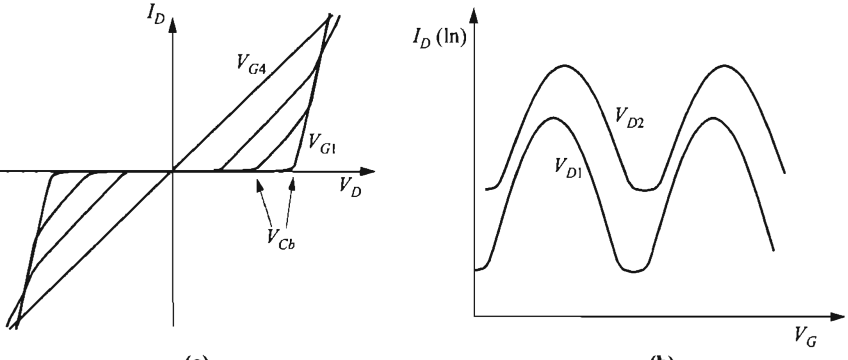


Fig. 54 (a) I-V characteristics of SET for various V_G. The Coulomb-blockade voltage can be varied by V_G. (b) Drain current (logarithmic scale) as a function of V_G for various V_D. Note that V_G is shifted by V_D.

$$\frac{q^2}{2C_j} = qV_j . \tag{130}$$

So V_j has to reach $q/2C_j$ before an electron can tunnel. This threshold voltage is the basis for the Coulomb blockade. Alternatively, one can get the same answer by considering the charging energy (E_{ch}) for transferring N_i number of electrons,

$$\begin{aligned} E_{ch} &= \frac{(Q_o - N_i q)^2}{2C_j} - \frac{Q_o^2}{2C_j} \\ &= \frac{N_i^2 q^2}{2C_j} - N_i q V_j \end{aligned} \tag{131}$$

where Q_o is the original charge before tunneling and is equal to V_j/C_j. The criterion for switching to a different state is that E_{ch} has to be negative and at minimum.

Next, we consider a single-electron box where an island is placed between two capacitors, the same as the situation when the source and drain of an SET is tied together (Fig. 53c). As the gate voltage is increased, the island voltage (V_I) is also increased accordingly, although scaled down by a factor of C_G/C_Σ. Similar to the case above, as soon as the tunneling junction gets above a voltage of $q/2C_\Sigma$, one electron starts to tunnel across it. Once an electron has tunneled to the center island, its potential drops by q/C_Σ. Figure 55 shows the charging of the center island and its potential as a function of the gate voltage. It can be seen that the gate voltage at which multiple values of N_i can coexist is at

$$V_G = \frac{q}{C_G}\left(N_i + \frac{1}{2}\right) . \tag{132}$$

This condition implies degeneracy: multiple N_i can exist without a change of energy, and one electron can tunnel freely to and from the island. One can imagine that for an

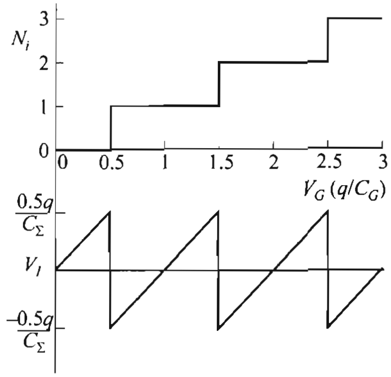


Fig. 55 Charging of single-electron box as a function of V_G (in q/C_G), and the corresponding island voltage V_I.

SET, if a small bias is applied to the drain, electrons can tunnel from the source to the island freely and subsequently from the island to the drain. This corresponds to the condition of V_G where the Coulomb blockade disappears in an SET.

Alternatively, one can derive Eq. 132 by considering the charging energy of a single-electron box,

$$E_{ch} = \frac{N_i^2 q^2}{2C_\Sigma} - \frac{N_i q V_G C_G}{C_\Sigma}. \tag{133}$$

By equating $E_{ch}(N_i + 1) = E_{ch}(N_i)$, the condition of Eq. 132 can be reached. Another approach to understand this is to plot E_{ch} vs. charge ($N_i q$) for different V_G, as shown in Fig. 56. Remembering that N_i takes on only integer values, there are only certain values of V_G where the E_{ch} minimum takes on two values of N_i, a condition of degeneracy. This means that the system can switch between these two states easily without any energy barrier.

We can now return to the SET and explain the two most-important phenomena: the Coulomb blockade and its voltage, and the Coulomb-blockade oscillations. For current to conduct from the source to drain, there are two junctions that electrons have

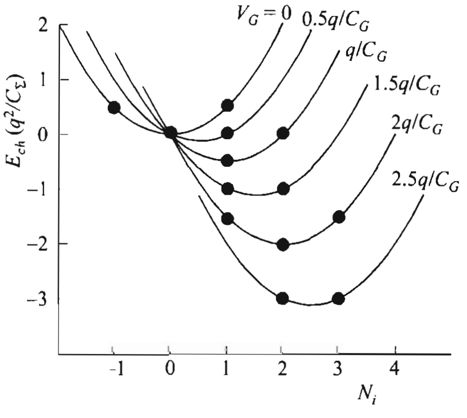


Fig. 56 E_{ch} vs. N_i for different V_G to determine N_i in single-electron box. It can be seen that depending on V_G, the E_{ch} minimum falls on either single or double values of N_i.

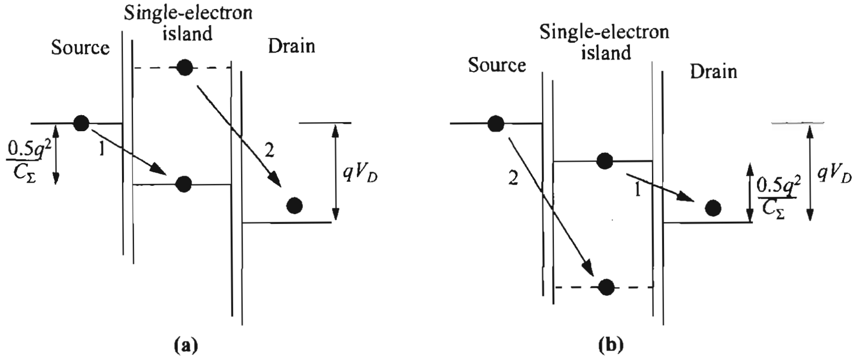

Single-electron island
Source
Drain
$\frac{0.5q^2}{C_\Sigma}$
1
2
qV_D
(a)
Single-electron island
Source
Drain
2
1
$\frac{0.5q^2}{C_\Sigma}$
qV_D
(b)

Fig. 57 Energy-band diagrams showing the sequence of tunneling events when the triggering process occurs at the junction (a) between the single-electron island and the source, and (b) between the island and the drain. Event-1 occurs before event-2. Note that the island potential changes by q/C_Σ after each tunneling event.

to tunnel through, but there is only one junction that controls the current flow. Using the energy-band diagrams of Fig. 57, if the bottleneck is at the junction between the source and the single-electron island, an electron will start to tunnel if that junction voltage exceeds $q/2C_\Sigma$, corresponding to the criterion of

$$\frac{V_D C_D}{C_\Sigma} + \frac{V_G C_G}{C_\Sigma} \geq \frac{q}{2C_\Sigma}, \tag{134}$$

giving a minimum value of V_D, or

$$V_{Cb} = \frac{q}{2C_D} - \frac{V_G C_G}{C_D}. \tag{135}$$

This blockade voltage as a function of V_G is shown in Fig. 58 as the line with a negative slope of

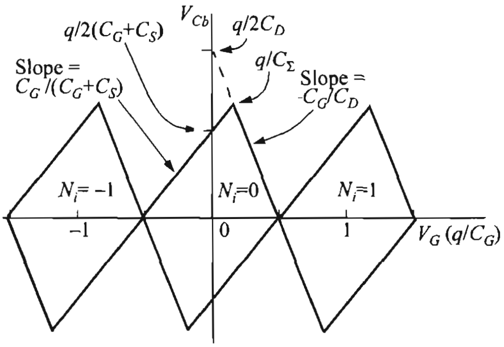


Fig. 58 Coulomb-blockade voltage V_{Cb} as a function of V_G forming the Coulomb-blockade diamonds.

$$\frac{dV_{Cb}}{dV_G} = -\frac{C_G}{C_D}. \tag{136}$$

Conversely, if the tunneling process is initiated by the island-drain junction, electrons start to flow if

$$V_D - \left(\frac{V_D C_D}{C_\Sigma} + \frac{V_G C_G}{C_\Sigma}\right) \geq \frac{q}{2C_\Sigma}, \tag{137}$$

giving another expression of

$$V_{Cb} = \frac{(q/2) + V_G C_G}{C_G + C_S}, \tag{138}$$

with a positive slope

$$\frac{dV_{Cb}}{dV_G} = \frac{C_G}{C_G + C_S}. \tag{139}$$

Note that since the current contours follow closely the shape of the Coulomb-blockade diamonds,[87] an SET has both positive and negative transconductance, depending on the V_G range. This is a unique feature that is different from a regular transistor. From the two lines of Eqs. 135 and 138, an intercept of q/C_Σ can be obtained which is the maximum V_{Cb}.

Alternatively, the Coulomb-blockade voltage can be obtained by setting the charging energy negative, either for the island-source junction or the island-drain junction as the current-limiting junction:

$$E_{ch}(N_i{=}1) = \frac{q^2}{2C_\Sigma} - q\left(\frac{V_D C_D}{C_\Sigma} + \frac{V_G C_G}{C_\Sigma}\right) \leq 0, \tag{140}$$

$$E_{ch}(N_i{=}1) = \frac{q^2}{2C_\Sigma} - q\left[V_D - \left(\frac{V_D C_D}{C_\Sigma} + \frac{V_G C_G}{C_\Sigma}\right)\right] \leq 0. \tag{141}$$

These equations lead to the same conclusions as Eqs. 135 and 138, respectively.

Above and below V_{Cb}, the current of an SET has been found to be described adequately by the orthodox theory,[88] which states that the tunneling rate is given by

$$T = \frac{\Delta E_{ch}}{q^2 R_T[1 - \exp(-\Delta E_{ch}/kT)]}, \tag{142}$$

where R_T is the total of $(R_{TS} + R_{TD})$ and ΔE_{ch} is the change of charging energy for different N_i states. One of the drawbacks of the SET is that besides V_G, the drain has substantial control of the current as well. As shown in Fig. 54b, a drain bias causes a shift of V_G by the amount [88]

$$\Delta V_G = \frac{(C_G + C_S - C_D)V_D}{2C_G}. \tag{143}$$

Within the Coulomb-blockade regime, the V_G swing needed to change the current by one order of magnitude is calculated to be

$$\Delta V_G \approx (\ln 10)\left(\frac{C_\Sigma}{C_G}\frac{kT}{q}\right). \tag{144}$$

Similarly, the V_D swing needed for the same change is given by

$$\Delta V_D \approx (\ln 10)\left(\frac{2kT}{q}\right). \tag{145}$$

So to have a transistor that has more gate control than drain control requires that the ratio C_G/C_Σ be larger than 0.5.

For application, the SET can perform logic. Since it can possess both positive and negative transconductance, a complementary type of logic can be employed using only one type of device, operated at different regimes. However, the small current and transconductance due to the tunneling nature limits its practicality in real circuits with parasitics. Another problem associated with SETs is the extreme sensitivity to parasitic charge surrounding the single-electron island, which is difficult to control. One potential application of the SET is nonvolatile memory, whose structure is shown in Fig. 44b except that the floating gate is miniaturized to a single-electron island. (More accurately, this memory cell employs a single-electron box or single-electron charging, but does not contain an SET.) A small number of electrons is stored or emptied in the single-electron island to control the threshold voltage of the MOSFET. The advantage is that since the charge in the floating-gate island is small and discrete, the signal of threshold voltage is quantized and the memory has multiple values.

REFERENCES

1. J. E. Lilienfeld, "Method and Apparatus for Controlling Electric Currents," U.S. Patent 1,745,175. Filed 1926. Granted 1930.
2. J. E. Lilienfeld, "Amplifier for Electric Currents," U.S. Patent 1,877,140. Filed 1928. Granted 1932.
3. J. E. Lilienfeld, "Device for Controlling Electric Current," U.S. Patent 1,900,018. Filed 1928. Granted 1933.
4. 0. Heil, "Improvements in or Relating to Electrical Amplifiers and other Control Arrangements and Devices," British Patent 439,457. Filed and granted 1935.
5. W. Shockley and G. L. Pearson, "Modulation of Conductance of Thin Films of Semiconductors by Surface Charges," *Phys. Rev.*, **74**, 232 (1948).
6. J. R. Ligenza and W. G. Spitzer, "The Mechanisms for Silicon Oxidation in Steam and Oxygen," *J. Phys. Chem. Solids*, **14**, 131 (1960).
7. M. M. Atalla. "Semiconductor Devices Having Dielectric Coatings," U.S. Patent 3,206,670. Filed 1960. Granted 1965.
8. D. Kahng and M. M. Atalla, "Silicon-Silicon Dioxide Field Induced Surface Devices," *IRE-AIEE Solid-State Device Res. Conf.*, (Carnegie Inst. of Tech., Pittsburgh, PA), 1960.
9. D. Kahng, "A Historical Perspective on the Development of MOS Transistors and Related Devices," *IEEE Trans. Electron Dev.*, **ED-23**, 655 (1976).

10. C. T. Sah, "Evolution of the MOS Transistor—From Conception to VLSI," *Proc. IEEE*, **76**, 1280 (1988).
11. H. K. J. Ihantola and J. L. Moll, "Design Theory of a Surface Field-Effect Transistor," *Solid-State Electron.*, **7**, 423 (1964).
12. C. T. Sah, "Characteristics of the Metal-Oxide-Semiconductor Transistors," *IEEE Trans. Electron Dev.*, **ED-11**, 324 (1964).
13. S. R. Hofstein and F. P. Heiman, "The Silicon Insulated-Gate Field-Effect Transistor," *Proc. IEEE*, **51**, 1190 (1963).
14. J. R. Brews, "Physics of the MOS Transistor," in D. Kahng, Ed., *Applied Solid State Science*, Suppl. 2A, Academic, New York, 1981.
15. Y. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd Ed., Oxford University Press, Oxford, 1999.
16. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, Cambridge, 1998.
17. R. M. Warner, Jr. and B. L. Grung, *MOSFET Theory and Design*, Oxford University Press, Oxford, 1999.
18. L. L. Chang and H. N. Yu, "The Germanium Insulated-Gate Field-Effect Transistor (FET)," *Proc. IEEE*, **53**, 316 (1965).
19. P. D. Ye, G. D. Wilk, J. Kwo, B. Yang, H. J. L. Gossmann, M. Frei, S. N. G. Chu, J. P. Mannaerts, M. Sergent, M. Hong, K. K. Ng, and J. Bude, "GaAs MOSFET with Oxide Gate Dielectric Grown by Atomic Layer Deposition," *IEEE Electron Dev. Lett.*, **EDL-24**, 209, (2003).
20. H. C. Pao and C. T. Sah, "Effects of Diffusion Current on Characteristics of Metal-Oxide (Insulator)-Semiconductor Transistors (MOST)," *IEEE Trans. Electron Dev.*, **ED-12**, 139 (1965).
21. A. S. Grove and D. J. Fitzgerald, "Surface Effects on *p-n* Junctions: Characteristics of Surface Space-Charge Regions under Nonequilibrium Conditions," *Solid-State Electron.*, **9**, 783 (1966).
22. J. R. Brews, "A Charge-Sheet Model of the MOSFET," *Solid-State Electron.*, **21**, 345 (1978).
23. D. M. Caughey and R. E. Thomas, "Carrier Mobilities in Silicon Empirically Related to Doping and Field," *Proc. IEEE*, **55**, 2192 (1967).
24. K. Natori, "Ballistic Metal-Oxide-Semiconductor Field Effect Transistor," *J. Appl. Phys.*, **76**, 4879 (1994).
25. K. Natori, "Scaling Limit of the MOS Transistor—A Ballistic MOSFET," *IEICE Trans. Electron.*, **E84-C**, 1029 (2001).
26. M. Lundstrom, "Elementary Scattering Theory of the Si MOSFET," *IEEE Electron Dev. Lett.*, **EDL-18**, 361 (1997).
27. F. Assad, Z. Ren, D. Vasileska, S. Datta, and M. Lundstrom, "On the Performance Limits for Si MOSFET's: A Theoretical Study," *IEEE Trans. Electron Dev.*, **ED-47**, 232 (2000).
28. M. Lundstrom, "Essential Physics of Carrier Transport in Nanoscale MOSFETs," *IEEE Trans. Electron Dev.*, **ED-49**, 133 (2002).
29. M. B. Barron, "Low Level Currents in Insulated Gate Field Effect Transistors," *Solid-State Electron.*, **15**, 293 (1972).

30. W. M. Gosney, "Subthreshold Drain Leakage Current in MOS Field-Effect Transistors," *IEEE Trans. Electron Dev.*, **ED-19**, 213 (1972).
31. G. W. Taylor, "Subthreshold Conduction in MOSFET's," *IEEE Trans. Electron Dev.*, **ED-25**, 337 (1978).
32. A. G. Sabnis and J. T. Clemens, "Characterization of the Electron Mobility in the Inverted ⟨100⟩ Si Surface," *Tech. Dig. IEEE IEDM*, p.18, 1979.
33. S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the Universality of Inversion Layer Mobility in Si MOSFET's: Part I—Effects of Substrate Impurity Concentration," *IEEE Trans. Electron Dev.*, **ED-41**, 2357 (1994).
34. J. A. Cooper, Jr. and D. F. Nelson, "High-Field Drift Velocity of Electrons at the Si-SiO_2 Interface as Determined by a Time-of-Flight Technique," *J. Appl. Phys.*, **54**, 1445 (1983).
35. L. Vadasz and A. S. Grove, "Temperature Dependence of MOS Transistor Characteristics Below Saturation," *IEEE Trans. Electron Dev.*, **ED-13**, 863 (1966).
36. F. Gaensslen, V. L. Rideout, E. J. Walker, and J. J. Walker, "Very Small MOSFET's for Low-Temperature Operation," *IEEE Trans. Electron Dev.*, **ED-24**, 218 (1977).
37. G. Merckel, "Ion Implanted MOS Transistors—Depletion Mode Devices," in F. Van de Wiele, W. L. Engle, and P. G. Jespers, Eds., *Process and Device Modeling for IC Design*, Noordhoff, Leyden, 1977.
38. J. S. T. Huang and G. W. Taylor, "Modeling of an Ion-Implanted Silicon-Gate Depletion-Mode IGFET," *IEEE Trans. Electron Dev.*, **ED-22**, 995 (1975).
39. T. E. Hendrikson, "A Simplified Model for Subpinchoff Condition in Depletion Mode IGFET's," *IEEE Trans. Electron Dev.*, **ED-25**, 435 (1978).
40. M. J. van der Tol and S. G. Chamberlain, "Potential and Electron Distribution Model for the Buried-Channel MOSFET," *IEEE Trans. Electron Dev.*, **ED-36**, 670 (1989).
41. R. H. Dennard, F. H. Gaensslen, H. Yu, V. L. Rideout, E. Bassons, and A. R. LeBlanc, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE J. Solid State Circuits*, **SC-9**, 256 (1974).
42. P. K. Chatterjee, W. R. Hunter, T. C. Holloway, and Y. T. Lin, "The Impact of Scaling Laws on the Choice of *n*-channel or *p*-channel for MOS VLSI," *IEEE Electron Dev. Lett.*, **EDL-1**, 220 (1980).
43. J. Meindl, "Circuit Scaling Limits for Ultra Large Scale Integration," *Digest Int. Solid-State Circuits Conf.*, 36, Feb. 1981.
44. G. Baccarani, M. R. Wordeman, and R. H. Dennard, "Generalized Scaling Theory and its Application to a 1/4 Micrometer MOSFET Design," *IEEE Trans. Electron Dev.*, **ED-31**, 452 (1984).
45. J. R. Brews, W. Fichtner, E. H. Nicollian, and S. M. Sze, "Generalized Guide for MOSFET Miniaturization," *IEEE Electron Dev. Lett.*, **EDL-1**, 2 (1980).
46. K. K. Ng, S. A. Eshraghi, and T. D. Stanik, "An Improved Generalized Guide for MOSFET Scaling," *IEEE Trans. Electron Dev.*, **ED-40**, 1895 (1993).
47. L. D. Yau, "A Simple Theory to Predict the Threshold Voltage of Short-Channel IGFET's," *Solid-State Electron.*, **17**, 1059 (1974).
48. W. Fichtner and H. W. Potzl, "MOS Modeling by Analytical Approximations. I. Subthreshold Current and Threshold Voltage," *Int. J. Electron.*, **46**, 33 (1979).

49. Y. Taur, G. J. Hu, R. H. Dennard, L. M. Terman, C. Y. Ting, and K. E. Petrillo, "A Self-Aligned 1 μm Channel CMOS Technology with Retrograde *n*-well and Thin Epitaxy," *IEEE Trans. Electron Dev.*, **ED-32**, 203 (1985).
50. W. Fichtner, "Scaling Calculation for MOSFET's," *IEEE Solid State Circuits and Technology Workshop on Scaling and Microlithography*, New York, Apr. 22, 1980.
51. K. K. Ng and G. W. Taylor, "Effects of Hot-Carrier Trapping in *n*- and *p*-Channel MOSFET's," *IEEE Trans. Electron Dev.*, **ED-30**, 871 (1983).
52. T. H. Ning, C. M. Osburn, and H. N. Yu, "Effect of Electron Trapping on IGFET Characteristics," *J. Electron. Mater.*, **6**, 65 (1977).
53. E. H. Nicollian and C. N. Berglund, "Avalanche Injection of Electrons into Insulating SiO_2 Using MOS Structures," *J. Appl. Phys.*, **41**, 3052 (1970).
54. T. Kamata, K. Tanabashi, and K. Kobayashi, "Substrate Current Due to Impact Ionization in MOSFET," *Jpn. J. Appl. Phys.*, **15**, 1127 (1976).
55. E. Sun, J. Moll, J. Berger, and B. Alders, "Breakdown Mechanism in Short-Channel MOS Transistors," *Tech. Dig. IEEE IEDM*, p. 478, 1978.
56. T. Y. Chan, A. T. Wu, P. K. Ko, and C. Hu, "Effects of the Gate-to-Drain/Source Overlap on MOSFET Characteristics," *IEEE Electron Dev. Lett.*, **EDL-8**, 326 (1987).
57. D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H. P. Wong, "Device Scaling Limits of Si MOSFETs and Their Application Dependencies," *Proc. IEEE*, **89**, 259 (2001).
58. B. Yu, H. Wang, A. Joshi, Q. Xiang, E. Ibok, M. Lin, "15nm Gate Length Planar CMOS Transistor," *Tech. Dig. IEEE IEDM*, p.937, 2001.
59. A. Hokazono, K. Ohuchi, M. Takayanagi, Y. Watanabe, S. Magoshi, Y. Kato, T. Shimizu, S. Mori, H. Oguma, T. Sasaki, et al., "14 nm Gate Length CMOSFETs Utilizing Low Thermal Budget Process with Poly-SiGe and Ni Salicide," *Tech. Dig. IEEE IEDM*, p.639, 2002.
60. K. K. Ng and W. T. Lynch, "Analysis of the Gate-Voltage-Dependent Series Resistance of MOSFETs," *IEEE Trans. Electron Dev.*, **ED-33**, 965 (1986).
61. M. P. Lepselter and S. M. Sze, "SB-IGFET: An Insulated-Gate Field-Effect Transistor Using Schottky Barrier Contacts as Source and Drain," *Proc. IEEE*, **56**, 1088 (1968).
62. R. R. Troutman, *Latchup in CMOS Technology: The Problem and its Cure*, Kluwer, Norwell, Massachusetts, 1986.
63. J. Kedzierski, P. Xuan, E. H. Anderson, J. Bokor, T. J. King, and C. Hu, "Complementary Silicide Source/Drain Thin-Body MOSFETs for the 20nm Gate Length Regime," *Tech. Dig. IEEE IEDM*, p.57, 2000.
64. S. Nishimatsu, Y. Kawamoto, H. Masuda, R. Hori, and O. Minato, "Grooved Gate MOSFET," *Jpn. J. Appl. Phys.*, **16**; Suppl. **16-1**, 179 (1977).
65. G. K. Celler and S. Cristoloveanu, "Frontiers of Silicon-on-Insulator," *J. Appl. Phys.*, **93**, 1 (2003).
66. K. A. Jenkins, J. Y. C. Sun, and J. Gautier, "History Dependence of Output Characteristics of Silicon-on-Insulator (SOI) MOSFET's," *IEEE Electron Dev. Lett.*, **EDL-17**, 7 (1996).
67. C. R. Kagan and P. Andry, Eds., *Thin-Film Transistors*, Marcel Dekker, New York, 2003.
68. D. Hisamoto, T. Kaga, and E. Takeda, "Impact of the Vertical 'DELTA' Structure on Planar Device Technology," *IEEE Trans. Electron Dev.*, **ED-38**, 1399 (1991).

69. B. S. Doyle, S. Datta, M. Doczy, S. Hareland, B. Jin, J. Kavalieros, T. Linton, A. Murthy, R. Rios, and R. Chau, "High Performance Fully-Depleted Tri-Gate CMOS Transistors," *IEEE Electron Dev. Lett.*, **EDL-24**, 263 (2003).
70. J. M. Hergenrother, G. D. Wilk, T. Nigam, F. P. Klemens, D. Monroe, P. J. Silverman, T. W. Sorsch, B. Busch, M. L. Green, M. R. Baker, et. al., "50 nm Vertical Replacement-Gate (VRG) nMOSFETs with ALD HfO_2 and Al_2O_3 Gate Dielectrics," *Tech. Dig. IEEE IEDM*, p.51, 2001.
71. T. Masuhara and R. S. Muller, "Analytical Technique for the Design of DMOS Transistors," *Jpn. J. Appl. Phys.*, **16**, 173 (1976).
72. M. D. Pocha, A. G. Gonzalez, and R. W. Dutton, "Threshold Voltage Controllability in Double-Diffused MOS Transistors," *IEEE Trans. Electron Dev.*, **ED-21**, 778 (1974).
73. A. W. Ludikhuize, "A Review of RESURF Technology," *Proc. 12th Int. Symp. Power Semiconductor Devices & ICs*, p.11, 2000.
74. P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, Eds., *Flash Memories*, Kluwer, Norwell, Massachusetts, 1999.
75. C. Hu, Ed., *Nonvolatile Semiconductor Memories: Technologies, Design, and Applications*, IEEE Press, Piscataway, New Jersey, 1991.
76. W. D. Brown and J. E. Brewer, Eds., *Nonvolatile Semiconductor Memory Technology*, IEEE Press, Piscataway, New Jersey, 1998.
77. D. Kahng and S. M. Sze, "A Floating Gate and Its Application to Memory Devices," *Bell Syst. Tech. J.*, **46**, 1283 (1967).
78. D. Frohman-Bentchkowsky, "FAMOS—A New Semiconductor Charge Storage Device," *Solid-State Electron.*, **17**, 517 (1974).
79. H. Iizuka, F. Masuoka, T. Sato, and M. Ishikawa, "Electrically Alterable Avalanche- Injection-Type MOS Read-Only Memory with Stacked-Gate Structures," *IEEE Trans. Electron Dev.*, **ED-23**, 379 (1976).
80. S. Mahapatra, S. Shukuri, and J. Bude, "CHISEL Flash EEPROM—Part I: Performance and Scaling," *IEEE Trans. Electron Dev.*, **ED-49**, 1296 (2002).
81. S. K. Lai and V. K. Dham, "VLSI Electrically Erasable Programmable Read Only Memory," in N. G. Einspruch, Ed., *VLSI handbook*, Academic Press, Orlando, FL, 1985.
82. Y. Nishi and H. Iizuka, "Nonvolatile Memories," in D. Kahng, Ed., *Applied Solid State Science*, Suppl. 2A, Academic, New York, 1981.
83. Y. Kamigaki and S. Minami, "MNOS Nonvolatile Semiconductor Memory Technology: Present and Future," *IEICE Trans. Electron.*, **E84-C**, 713 (2001).
84. D. V. Averin and K. K. Likharev, "Coulomb Blockade of Single-Electron Tunneling, and Coherent Oscillations in Small Tunnel Junctions," *J. Low Temp. Phys.*, **62**, 345 (1986).
85. T. A. Fulton and G. J. Dolan, "Observation of Single-Electron Charging Effects in Small Tunnel Junctions," *Phys. Rev. Lett.*, **59**, 109 (1987).
86. M. A. Kastner, "Artificial Atoms," *Physics Today*, 24 (Jan. 1993).
87. Y. A. Pashkin, Y. Nakamura and J. S. Tsai, "Room-Temperature Al Single-Electron Transistor Made by Electron-Beam Lithography," *Appl. Phys. Lett.*, **76**, 2256 (2000).
88. K. Uchida, K. Matsuzawa, J. Koga, R. Ohba, S. Takagi and A. Toriumi, "Analytical Single-Electron Transistor (SET) Model for Design and Analysis of Realistic SET Circuits," *Jpn. J. Appl. Phys.*, **39**, 2321 (2000).

PROBLEMS

1. Derive Eq. 23 from Eqs. 20 and 22 (p.303).

2. For a square MOSFET ($Z/L = 1$), I_D is measured to be 18.7 μA at $V_D = 0.4$ V and $V_G = 3$ V. If we require a current of 1.6 mA at $V_D = 0.4$ V and $V_G = 3$ V, what is the minimum width Z of the device? Assume that the polysilicon gate length is 0.6 μm and the n^+ source and drain each diffuses sideways 0.05 μm under the gate.

3. Consider a submicron MOSFET with L = 0.25 μm, Z = 5 μm, $N_A = 10^{17}$ cm^{-3}, μ_n = 500 cm^2/V-s, $C_{ox} = 3.45\times10^{-7}$ F/cm^2, and $V_T = 0.5$ V, find the channel conductance for $V_G = 1$ V and $V_D = 0.1$ V.

4. For a MOSFET with a channel length of 10 μm under certain biasing conditions, the channel current I_D is 1 mA and the gate current is 1 μA. We want to reduce the gate current to 10 $^{6}I_D$ under the same biasing condition and for the same device parameters except the channel length. Find the channel length.

5. For an MOSFET with sufficient drain voltage to be in saturation (under constant-mobility condition), the current is 50 μA at V_G = 1 V, and 200 μA at V_G = 3 V. Find the threshold voltage.

6. (a) To avoid hot-electron effect in an n-channel MOSFET, we assume an allowed maximum field in the oxide to be 1.45×10^6 V/cm. Find the corresponding surface potential ψ_s in the silicon for a doping concentration of 10^{18} cm^{-3}.
   (b) For an n^+-polysilicon gate, find the threshold voltage of the above MOSFET with d = 8 nm, assuming $Q_{it} = Q_{ox} = Q_f = Q_m = 0$.

7. An n-channel MOSFET is designed to have a threshold voltage of +0.5 V and a gate oxide thickness of 15 nm. Find the channel doping to give the desired V_T if n^+-polysilicon is used as the gate material, and there are no oxide charge, interface-trapped charge, and mobile ions in the device.

8. To isolate devices from interacting with each other, each MOSFET is surrounded by a field oxide. If the "field transistor" associated with the field oxide must have a threshold voltage of ≥ 20V, calculate the minimum field-oxide thickness. $N_A = 10^{17}$ cm^{-3}, $Q_f/q = 10^{11}$ cm^{-3}, and an n^+-polysilicon is used for local interconnect as the gate electrode.

9. An n-channel n^+-polysilicon-SiO$_2$-Si MOSFET has $N_A = 10^{17}$ cm^{-3}, $Q_f/q = 2\times10^{10}$ cm^{-2}, and d = 10 nm. Boron ions are implanted to increase the threshold voltage to +1 V. Find the implant dose, assume that the implanted ions form a sheet of negative charge at the Si-SiO$_2$ interface. ($\phi_{ms} = -0.98$ V).

10. For an n-MOSFET with $q\psi_B$ of 0.5 eV, the threshold voltage change ΔV_T is 1 V when a substrate bias V_{BS} of −1 V is applied. What is ΔV_T when V_{BS} is −3 V?

11. A MOSFET ($N_A = 10^{17}$ cm^{-3}, d = 5 nm) has a threshold voltage of V_T = 0.5 V, a subthreshold swing of 100 mV/decade, and a drain current of 0.1 μA at V_T. If we want to reduce the leakage current at V_G = 0 to 10^{-13} A, find the reverse substrate-source bias required to achieve the reduction.

12. The subthreshold current of an ideal MOSFET is given by

$$I_D = A(\beta\psi_s)^{-1/2}\exp(\beta\psi_s),$$

$$\beta\psi_s = \beta V_G - \frac{a^2}{2\beta}\left[\sqrt{1 + \frac{4}{a^2}(\beta V_G - 1)} - 1\right],$$

where ψ_s is the surface potential, $\beta \equiv q/kT$, $a \equiv \sqrt{2}\,(\varepsilon_s/\varepsilon_{ox})(t_{ox}/L_D)$, L_D is Debye length $= \sqrt{\varepsilon_s/qN_A\beta}$, and A = constant. Show that the subthreshold swing S is given by

$$S \equiv (\ln 10)\frac{dV_G}{d(\ln I_D)} = \frac{kT}{q}(\ln 10)\left(1 + \frac{C_D}{C_{ox}}\right)$$

where $C_D \equiv \sqrt{q\varepsilon_s N_A/2\psi_s}$, $C_{ox} \equiv \varepsilon_{ox}/t_{ox}$, and $a \gg C_D/C_{ox}$.

13. For a MOSFET with a gate oxide of 10 nm and a substrate doping of 10^{17} cm^{-3}, find the subthreshold swing.

14. For a Si MOSFET with $N_A = 5\times10^{16}$ cm^{-3}, d =10 nm, and an interface-trap density of 10^{11} cm^{-2}, find the subthreshold swing with a grounded substrate terminal.

15. An idealized implanted step doping profile has $N_S = 10^{16}$ cm^{-3}, $N_B = 10^{15}$ cm^{-3}, and x_s = 0.3 μm. Find (1) the implanted dose D_I, (2) the centroid of the dose, and (3) the threshold voltage shift (d = 100 nm) with respect to a uniformly doped N_B case.

16. Derive Eq. 79.

17. Refer to Fig. 21 (p. 322), assume $N_B = 7.5\times10^{15}$ cm^{-3}, d = 35 nm, a reverse back bias of 1 V, and an implanted dose $D_I = 6\times10^{11}$ cm^{-2}, find the depth of the centroid at which the depletion-layer edge is clamped to the implant (in nm).

18. Find the scaled drain current per unit channel width (I_D/Z) for two n-MOSFETs, one with constant-voltage scaling and one with constant-field scaling. Assume the devices are operated under velocity-saturation condition. The initial device parameters are L = 1 μm, d = 10 nm, V_D = 5 V, I_D/Z = 500 μA/μm. The scaling factor is $\kappa = 5$.

19. For the constant-voltage scaling approach of MOSFET with a scaling factor κ=10, find the doping concentration of a scaled device if the original device has a doping of 10^{15} cm^{-3} (in cm^{-3}).

20. When the linear dimensions of a MOSFET are scaled down by a factor of 10 based on the constant-field scaling, (a) find the corresponding factor of the scaled switching energy, and (b) find the scaled power-delay product, assuming the product is 1 J for the original large device.

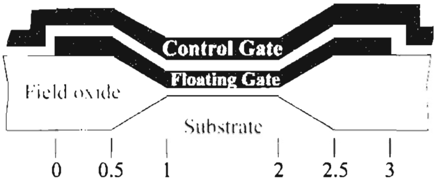
21. A composite structure of a 20 nm Ta_2O_5 ($\varepsilon_i/\varepsilon_o$ = 25) and a 2-nm SiO_2 is sandwiched between a top and bottom electrodes. Find the equivalent SiO_2 thickness (in nm).

22. A DRAM must operate with a minimum refresh time of 4 ms. The storage capacitor in each cell has a capacitance of 50 fF and is fully charged at 5 V. Estimate the worst-case leakage current that the dynamic capacitor node can tolerate (i.e., the charge in the capacitor has dropped to its 50% level).

23. For DRAM operation assume that we need a minimum of 10^5 electrons for the MOS storage capacitor. If the capacitor has an area of 0.5 μm×0.5 μm on the wafer surface, an oxide thickness of 5 nm, and is fully charged to 2 V, what is the required minimum depth of a rectangular-trench capacitor?

24. For a floating-gate nonvolatile memory device, the lower insulator has a dielectric constant of 4 and is 10 nm thick. The insulator above the floating gate has a dielectric constant of 10 and is 100 nm thick. If the current density $J_1 = \sigma\mathscr{E}_1$ where $\sigma = 10^{-7}$ S/cm, and the current

in the upper insulator is zero, find the threshold voltage shift for a sufficiently long time such that J_1 becomes negligibly small. The applied voltage on the control gate is 10 V.



25. Consider a NVSM cell whose cross section is shown. The channel width is 1 μm. Assume that the bird's beak has the linear wedge shape illustrated. The gate oxide thickness (between substrate and FG) is 35 nm, the interpoly dielectric is an oxide of 50 nm, and the field oxide is 0.6 μm. The physical gate length is 1.2 μm, the metallurgical junction lies 0.15 μm under the gate and the effective channel length is 0.7 μm. The floating gate poly is 0.3 μm thick. Calculate (a) the value of the control gate to floating gate capacitance, (b) the drain to floating gate capacitance, assume half of the channel capacitance to the source and other half to the drain, and (c) if the floating gate to substrate capacitance is 0.14 fF, calculate the control gate to floating gate coupling ration, R_{CG}, and the drain to floating gate coupling ration, R_D.

26. For a silicon nonvolatile memory with a floating gate, the thickness and the dielectric constant for the first insulator (thermally grown SiO_2) are 3 nm and 3.9, and the corresponding values for the second insulator are 30 nm and 30. Estimate the stored charge/cm^2 in the floating gate after a gate voltage of 5.52 V is applied for 1 ms. There is no current conduction through the second insulator, and the current in the first insulator is by Fowler-Nordheim tunneling.

27. A floating-gate nonvolatile semiconductor memory has a total capacitance of 3.71 fF, a control gate to floating gate capacitance of 2.59 fF, a drain to floating gate capacitance of 0.49 fF, a floating gate to substrate capacitance of 0.14 fF. How many electrons are needed to shift the measured threshold by 0.5 V (measured from the control gate)?

28. An EEPROM has $C_{CG} = 2.59$ fF, $C_S = C_D = 0.49$ fF, and $C_B = 0.14$ fF, where they indicate the capacitances between the floating gate and the control gate, source, drain, and substrate, respectively. Assume that when the control gate and the floating gate are shorted together, the device threshold is measured to be 1.5 V. If the control gate is at 12 V, and the drain is at 7 V during programming, to what potential can the floating gate be charged while the programming voltages are present? What threshold would be observed after programming under these biases for a drain bias during reading of 2 V?

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

7

JFETs, MESFETs, and MODFETs

7.1 INTRODUCTION
7.2 JFET AND MESFET
7.3 MODFET

7.1 INTRODUCTION

In this chapter, we discuss field-effect transistors (FETs) other than the MOSFET to which we have devoted the whole chapter. Referring back to the FET family tree depicted in Fig. 3 (p. 295) of Chapter 6, we pointed out that all FETs have a gate that is coupled to the channel through some form of a capacitor. While in a MOSFET the capacitor is formed by an oxide layer, the JFET (junction FET) and the MESFET (metal-semiconductor FET) form the capacitor by virtue of a depletion layer in a junction; the JFET from a *p-n* junction and the MESFET from a Schottky (metal-semiconductor) junction. In the branch of HFET (heterojunction FET), a layer of high-bandgap material is grown epitaxially over the channel, and it is used as an insulator. Bear in mind that the conductivity of a material is fundamentally related to its energy gap. An insulator is characterized by having a large energy gap. Epitaxial heterojunction produces an ideal interface. Such technique is necessary when an ideal oxide-semiconductor interface is lacking, which is practically for all semiconductors other than silicon. Under HFETs, the high-bandgap material can be doped or undoped. With high-E_g material that is doped, carriers from dopants are transferred to the heterointerface and form a channel of high mobility, since the channel itself is undoped to avoid impurity scattering. This technique is called *modulation doping*. When applied to the gate of an FET, the result is a MODFET (modulation-doped FET) which possesses some interesting features. When the high-E_g material is undoped, the resultant device is called HIGFET (heterojunction insulated-gate FET). In this case modulation doping is not present and the high-E_g material is used purely as an insulator. Such a device behaves in principle the same way as a MOSFET and will not be discussed further in this chapter. So the chapter primarily focuses on JFET, MESFET, and MODFET.

Of the three devices, the JFET and MESFET share a similar working principle. They are both based on bulk, buried-channel conduction. Their current path is modulated by the depletion width under the gate. They are also similar to a buried-channel MOSFET, except in the latter, the gate can be forward biased to the extend that accumulation at the surface can occur such that a surface channel can be formed in parallel to the buried channel. However in the JFET and MESFET, the junctions cannot be biased beyond or even close to flat-band before excessive current flows through the gate. So the JFET and MESFET are first studied together sharing the same equations. The MODFET has a two-dimensional channel at the heterointerface and will be treated independently.

7.2 JFET AND MESFET

The JFET, first proposed and analyzed by Shockley in 1952,[1] is basically a voltage-controlled resistor. Based on Shockley's theoretical treatment, the first working JFET was reported by Dacey and Ross, who later also considered the effect of field-dependent mobility.[2,3]

The MESFET was proposed and first demonstrated by Mead in 1966.[4] Shortly after, microwave performance was reported by Hooper and Lehrer in 1967, using a GaAs epitaxial layer on semiinsulating GaAs substrate.[5]

Both JFET and MESFET have the advantage of avoiding problems related to the oxide-semiconductor interface in a MOSFET, such as interface traps and reliability issues arising from hot-electron injection and trapping. However, they have limitation on bias range allowed on the input gate. In comparison, the MESFET offers certain processing and performance advantages over the JFET. The metal gate requires only low-temperature processing compared to a *p-n* junction made by diffusion or implant-anneal sequence. The low gate resistance and low *IR* drop along the channel width is a big factor in microwave performance such as noise and $f_{\max}$. The metal gate has better control in defining short channel lengths for high-speed applications. It can also serve as an efficient heat sink for power applications. On the other hand, the JFET has a more robust junction for higher breakdown and power capability. A *p-n* junction has a higher built-in potential which is useful towards achieving an enhancement-mode device. The higher potential also reduces the gate leakage for the same bias. The *p-n* junction is a more controlled structure whereas a good Schottky barrier sometimes is difficult to form on certain semiconductors such as some *p*-type materials. A JFET has more freedom for various gate configurations, such as a heterojunction or a buffered-layer gate, that improve certain aspects of performance.

7.2.1 *I-V* Characteristics

Similarity between JFET and MESFET can be seen from their schematic diagrams shown in Fig. 1, using an *n*-type channel as an example. The transistors consist of a

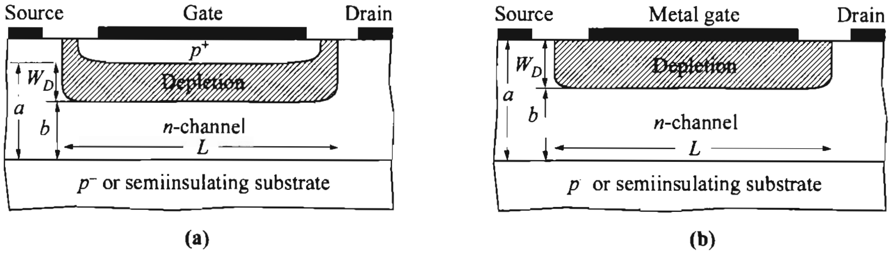


Fig. 1 Schematic structures of (a) JFET and (b) MESFET, showing their similarity in that the net channel opening b is controlled by the depletion width W_D.

conductive channel provided with two ohmic contacts, one acting as the source and the other as the drain. When a positive voltage V_D is applied to the drain with respect to the source, electrons flow from source to drain. Hence the source acts as the origin of the carriers and the drain as the sink. The third electrode, the gate, forms a rectifying junction and controls the net opening of the channel by varying the depletion width. The rectifying gate is a p-n junction in the JFET and is a Schottky-barrier junction in a MESFET. The device is basically a voltage-controlled resistor, and its resistance can be controlled by varying the width of the depletion layer extending into the channel region.

In Fig. 1 the basic device dimensions are the channel length L (also called the gate length), channel depth a, depletion-layer width W_D, net channel opening b, and channel width Z (into the paper, not shown). The voltage polarities shown are for an n-channel FET, and the polarities will be inverted for a p-channel FET. The gate and drain voltages are measured with respect to the source, and the source electrode is generally grounded. When $V_G = V_D = 0$, the device is in equilibrium and there is no current conduction. Most JFETs and MESFETs are of depletion mode, i.e., normally-on with $V_G = 0$, or threshold voltage V_T is negative. For a given V_G above the threshold voltage, the channel current increases with the drain voltage. Eventually for sufficiently large V_D, the current will saturate to a value $I_{D\text{sat}}$.

Very often for JFETs, the channel is surrounded by two gates. For Fig. 1a, that would be a second gate from the bottom side. The analysis we follow here is for one gate only. So this type of structure can be bisected into two halves and the final result becomes half of the total values, in both current and transconductance.

The basic current-voltage characteristics of a JFET or MESFET are shown in Fig. 2, where the drain current is plotted against the drain voltage for various fixed gate voltages. We can divide the characteristic into three regions: the linear region where the drain voltage is small and I_D is proportional to V_D; the nonlinear region; and the saturation region where the current remains essentially constant and is independent of V_D. As the gate bias becomes more negative, both the saturation current $I_{D\text{sat}}$ and the corresponding saturation voltage $V_{D\text{sat}}$ decrease. The locus of $I_{D\text{sat}}$-$V_{D\text{sat}}$ is shown in Fig. 2.

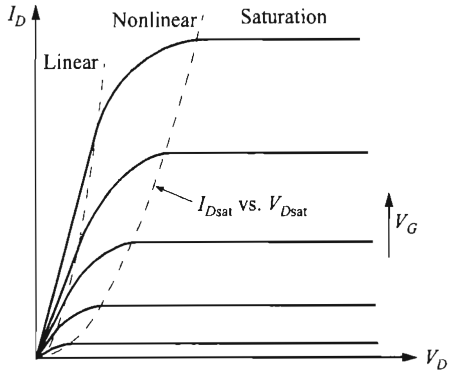


Fig. 2 General *I-V* characteristics of the JFET and MESFET.

We shall now derive the general *I-V* characteristics for JFETs and MESFETs, based on the following assumptions: (1) uniform channel doping, (2) gradual-channel approximation ($\mathscr{E}_x \ll \mathscr{E}_y$), (3) abrupt depletion layer, and (4) negligible gate current. We start with the channel charge distribution which is related to the channel dimensions. The channel dimensions and its potential distributions under both gate and drain biases are shown in more details in Fig. 3. These are the basis for deriving the *I-V* characteristics.

Channel-Charge Distribution. For a uniformly doped *n*-channel, under the gradual-channel approximation, the depletion-layer width W_D varies only gradually along the channel (*x*-direction), and one can solve the one-dimensional Poisson equation in the *y*-direction;

$$\frac{d\mathscr{E}_y}{dy} = -\frac{d^2\Delta\psi_i}{dy^2} = \frac{qN_D}{\varepsilon_s}, \tag{1}$$

where $\mathscr{E}_y$ is the electric field in the *y*-direction. The depletion-layer width at any distance *x* from the source is given by the one-sided abrupt-junction depletion approximation

$$W_D(x) = \sqrt{\frac{2\varepsilon_s[\psi_{bi} + \Delta\psi_i(x) - V_G]}{qN_D}}, \tag{2}$$

where ψ_{bi} is the built-in potential. For a JFET, the ψ_{bi} is that of a p^+-n junction, given by

$$\psi_{bi} \approx \frac{1}{q}\left[E_g - kT\ln\left(\frac{N_C}{N_D}\right)\right]. \tag{3}$$

For a MESFET, the ψ_{bi} is determined by the Schottky barrier height ϕ_{Bn} of the metal-semiconductor junction, given by

$$\psi_{bi} = \phi_{Bn} - \frac{kT}{q}\ln\left(\frac{N_C}{N_D}\right). \tag{4}$$

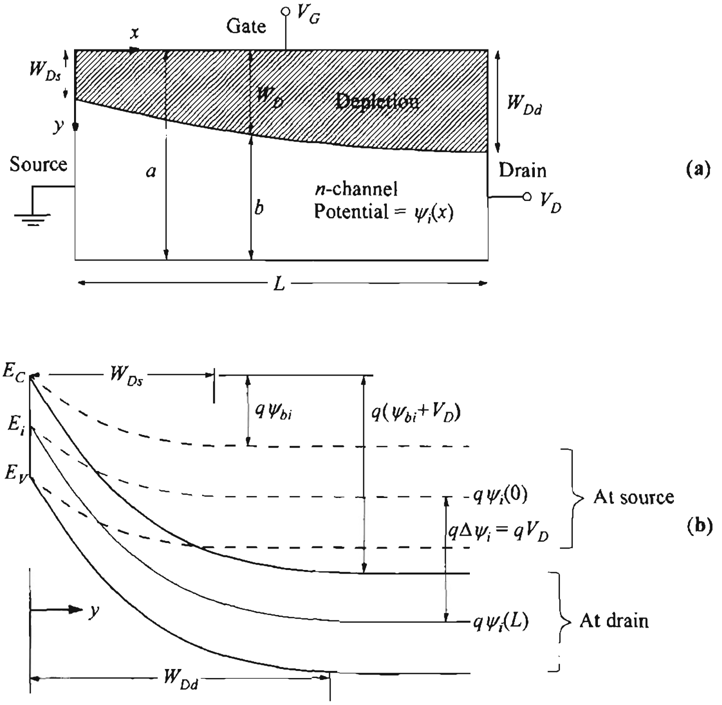

Gate
V_G
x
W_{Ds}
y
W_D
Depletion
W_{Dd}
Source
a
b
n-channel
Potential = $\psi_i(x)$
Drain
V_D
(a)
L
E_C
E_i
E_V
W_{Ds}
$q\psi_{bi}$
$q(\psi_{bi}+V_D)$
$-q\psi_i(0)$
$q\Delta\psi_i = qV_D$
At source
(b)
y
$-q\psi_i(L)$
At drain
W_{Dd}

Fig. 3 (a) Channel dimensions under drain and gate biases. (b) Energy-band diagram in y-direction at the source end (dashed lines) and drain end (solid lines).

The potential difference $\Delta\psi_i(x)$ is the potential of the neutral channel $[-E_i(x)/q]$ with respect to the source. So at the drain end, $\Delta\psi_i(L) = V_D$. The depletion widths at the source and drain ends are given by

$$W_{Ds} = W_D(0) = \sqrt{\frac{2\varepsilon_s(\psi_{bi} - V_G)}{qN_D}}, \tag{5}$$

$$W_{Dd} = W_D(L) = \sqrt{\frac{2\varepsilon_s(\psi_{bi} + V_D - V_G)}{qN_D}}. \tag{6}$$

The maximum value of gate bias applied to increase the current is limited to $V_G = \psi_{bi}$ which corresponds to the condition of $W_{Ds} = 0$. This flat-band condition in practice is not achievable due to the excessive forward current of the gate junction. The maximum value of W_{Dd} is equal to a, and the corresponding total band bending is called the *pinch-off potential*, defined as

$$\psi_P \equiv \frac{qN_Da^2}{2\varepsilon_s}. \tag{7}$$

The channel charge density, which is responsible for the current conduction, is proportional to the net channel opening, given by

$$Q_n(x) = qN_D(a - W_D). \tag{8}$$

The channel current is simply given by the charge multiplied by its velocity v,

$$I_D(x) = ZQ_n(x)v(x). \tag{9}$$

Since the current has to be continuous throughout the channel, it is independent of position. Integrating Eq. 9 from source to drain yields

$$I_D = \frac{Z}{L}\int_0^L Q_n(x)v(x)dx. \tag{10}$$

This is the basic equation used to derive the I-V relationship.

Equation 10 requires knowledge of the carrier velocity under an applied field, so the v-$\mathscr{E}$ relationship is critical. In the following analysis, we use different assumptions for such relationship. Referring again to Fig. 2, we find that current saturation can originate from two very different mechanisms. The first is due to channel pinch-off when the net channel is totally pinched off by the depletion width. This is called long-channel behavior, and found to be modelled well simply by a constant mobility, i.e. $v = \mu\mathscr{E}$. The second possible mechanism, especially true for short-channel devices, is that the field is high enough such that mobility is no longer constant and eventual the velocity rises to a constant value called saturation velocity. This occurs before the channel is pinched off. These effects are considered in the following subsections.

Constant Mobility. With constant mobility, the relationship $v = \mu\mathscr{E}_x$ is assumed to hold without limit. Using this relationship and that $\mathscr{E}_x = d\Delta\psi_i/dx$, Eq. 10 after carrying out the integration gives

$$\begin{aligned} I_D &= \frac{Zq\mu N_D}{L}\int_0^{V_D}\left[a - \sqrt{\frac{2\varepsilon_s(\psi_{bi} + \Delta\psi_i - V_G)}{qN_D}}\right]d\Delta\psi_i \\ &= G_i\left\{V_D - \frac{2}{3\sqrt{\psi_P}}[(\psi_{bi} + V_D - V_G)^{3/2} - (\psi_{bi} - V_G)^{3/2}]\right\}, \end{aligned} \tag{11}$$

where

$$G_i \equiv \frac{Zq\mu N_D a}{L} \tag{12}$$

is the full channel conductance when $W_D = 0$.

In the linear region, $V_D \ll V_G$ and $V_D \ll \psi_{bi}$, Eq. 11 is reduced to

$$I_{D\text{lin}} = G_i\left(1 - \sqrt{\frac{\psi_{bi} - V_G}{\psi_P}}\right)V_D \tag{13}$$

where ohmic characteristics are observed. Equation 13 can be further simplified by Taylor's expansion around $V_G = V_T$ to

$$I_{D\text{lin}} \approx \frac{G_i}{2\psi_P}(V_G - V_T)V_D \qquad V_G \approx V_T \tag{14}$$

with

$$V_T = \psi_{bi} - \psi_P \ . \tag{15}$$

V_T is the gate threshold voltage around which the transistor is turned on and off.

When the drain bias continues to increase, the current according to Eq. 11 goes through the nonlinear region. It reaches a peak and actually drops beyond that point. The drop of current is not physical, but it corresponds to a pinch-off condition when $W_{Dd} = a$. The V_D at the onset of this condition can be shown to occur at

$$V_{D\text{sat}} = \psi_P - \psi_{bi} + V_G = V_G - V_T \ . \tag{16}$$

Once that is known, the current in the saturation region can be found by substituting $V_{D\text{sat}}$ into Eq. 11:

$$I_{D\text{sat}} = G_i\left[\frac{\psi_P}{3} - (\psi_{bi} - V_G)\left(1 - \frac{2}{3}\sqrt{\frac{\psi_{bi} - V_G}{\psi_P}}\right)\right] . \tag{17}$$

It is seen here that $I_{D\text{sat}}$ is limited to a maximum of $G_i\psi_P/3$, a condition that cannot be reached in reality due to excessive gate current. The transconductance is given by

$$g_m \equiv \frac{dI_{D\text{sat}}}{dV_G} = G_i\left(1 - \sqrt{\frac{\psi_{bi} - V_G}{\psi_P}}\right) . \tag{18}$$

Qualitatively, for drain bias higher than $V_{D\text{sat}}$, the pinch-off point starts to migrate toward the source. However, the potential at the pinch-off point remains to be $V_{D\text{sat}}$, independent of V_D. The field within the drift region thus remains fairly constant, giving rise to current saturation. Practical devices show that $I_{D\text{sat}}$ does not saturate completely with V_D. This is due to the reduction in the effective channel length, which is measured between the source and the pinch-off point. Equation 17 can also be simplified using Taylor's expansion around $V_G = V_T$:

$$I_{D\text{sat}} \approx \frac{G_i}{4\psi_P}(V_G - V_T)^2 \qquad V_G \approx V_T , \tag{19}$$

and

$$g_m \approx \frac{G_i}{2\psi_P}(V_G - V_T) \qquad V_G \approx V_T . \tag{20}$$

It is seen here that the forms of Eqs. 14, 19, and 20 are similar to that of MOSFET only near the threshold, i.e., $V_G \approx V_T$. This stems from the fact that the gate capacitance (or depletion width) is gate-bias dependent in JFET and MESFET, while that in the MOSFET (gate dielectric) is fixed. In other words, in a MOSFET the channel charge is linearly dependent on V_G, while that is not true for a JFET or MESFET (Eq. 8).

One major difference of a bulk three-dimensional channel, as in JFET and MESFET, from a charge-sheet two-dimensional channel, as in MOSFET and MODFET, is that the current is controlled by the net channel opening. Because of this, it is possible that the current is expressed in terms of physical dimensions. This can offer a look from another angle and perhaps helps to understand the problem. Using the relationship

$$\frac{dW_D}{d\Delta\psi_i} = \frac{\varepsilon_s}{qN_D W_D}, \tag{21}$$

Equation 10 leads to

$$\begin{aligned} I_D &= \frac{Z\mu q^2 N_D^2}{\varepsilon_s L}\int_{W_{Ds}}^{W_{Dd}} (a - W_D) W_D dW_D \\ &= \frac{Z\mu q^2 N_D^2 a^3}{6\varepsilon_s L}[3(u_d^2 - u_s^2) - 2(u_d^3 - u_s^3)], \end{aligned} \tag{22}$$

where the normalized dimensionless units are defined as

$$u_d \equiv \frac{W_{Dd}}{a} = \sqrt{\frac{\psi_{bi} + V_D - V_G}{\psi_P}}, \tag{23}$$

$$u_s \equiv \frac{W_{Ds}}{a} = \sqrt{\frac{\psi_{bi} - V_G}{\psi_P}}. \tag{24}$$

Equation 22 can also be transformed directly from Eq. 11. In the linear region for small V_D, this equation can be shown to reduce to

$$I_{D\mathrm{lin}} = G_i(1 - u_s)V_D. \tag{25}$$

Current saturation is determined when the channel is pinched off. Setting $u_d = 1$, the saturation current is given by

$$I_{D\mathrm{sat}} = \frac{Z\mu q^2 N_D^2 a^3}{6\varepsilon_s L}(1 - 3u_s^2 + 2u_s^3). \tag{26}$$

Consequently, the transconductance is given by

$$\begin{aligned} g_m &= \frac{dI_{D\mathrm{sat}}}{dV_G} = \frac{dI_{D\mathrm{sat}}}{du_s} \times \frac{du_s}{dV_G} \\ &= G_i(1 - u_s)\ . \end{aligned} \tag{27}$$

Velocity-Field Relationship. For long-channel devices, the field is low enough that the carrier velocity is treated as being proportional to the field, i.e., constant mobility. For FETs with short channels, significant discrepancies are encountered between experiment and basic theory. One main reason for the discrepancies is the higher internal field for short channels. Figure 4 shows the qualitative dependence of the drift velocity versus electric field for silicon. At low fields the drift velocity increases linearly with the field, and the slope corresponds to a constant mobility ($\mu = v/\mathscr{E}$). At

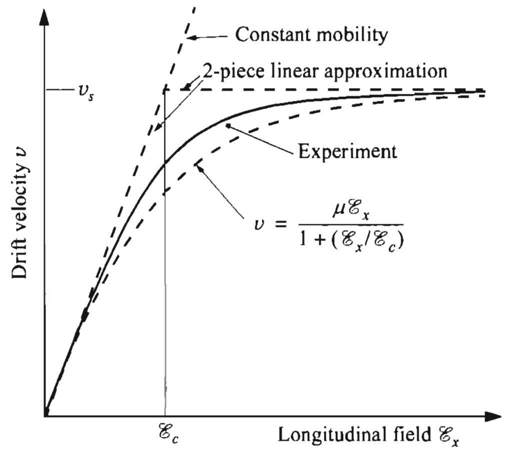


Fig. 4 Drift velocity vs. electric field for Si and semiconductors that do not have a transferred-electron effect.

higher fields, the carrier velocity deviates from a linear dependence. It becomes lower than simple extrapolation from the low-field slope, and eventually saturates to a value called saturation velocity v_s. So for short-channel devices, these effects have to be taken into account.

For silicon the drift velocity approaches its saturation value of 10^7 cm/s at fields above 5×10^4 V/cm. For some semiconductors such as GaAs and InP, the drift velocity first reaches a peak value and then decreases toward a saturation velocity of about 6–8×10^6 cm/s. This negative-resistance phenomenon is due to the transferred-electron effect. Its v-$\mathscr{E}$ relationship is too complicated to yield an analytical result and is not considered in this chapter.

In this section, we will examine two simple v-$\mathscr{E}$ relationships. The first is the two-piece linear approximation shown in Fig. 4. The second is an empirical formula which has a smooth transition between the constant-mobility regime to the saturation-velocity regime, given as

$$v(\mathscr{E}_x) = \frac{\mu\mathscr{E}_x}{1+(\mu\mathscr{E}_x/v_s)} = \frac{\mu\mathscr{E}_x}{1+(\mathscr{E}_x/\mathscr{E}_c)}, \tag{28}$$

where $\mathscr{E}_x = d\Delta\psi_i/dx$ is the longitudinal field in the channel. As seen, both relationships contain an important parameter, the critical field $\mathscr{E}_c$.

Field-Dependent Mobility: Two-Piece Linear Approximation. We first discuss velocity saturation based on the two-piece linear approximation. Note that in this model, the constant-mobility results (i.e., Eq. 11) are valid up to the point where the maximum field, which is at the drain end, reaches the critical field $\mathscr{E}_c$. Once at that $V_{D\text{sat}}$, which is lower than the $V_{D\text{sat}}$ of the constant-mobility model, the current saturates at a new and lowered $I_{D\text{sat}}$. So the main task is to calculate this new $V_{D\text{sat}}$. We start with Eq. 9 (and substitute $v = \mu\mathscr{E}$) which contains the relationship between field and current. Setting $\mathscr{E} = \mathscr{E}_c$ and $I_D = I_{D\text{sat}}$, we have

$$I_{D\mathrm{sat}} = Zq\mu N_D\mathscr{E}_c\left[a - \sqrt{\frac{2\varepsilon_s(\psi_{bi} + V_{D\mathrm{sat}} - V_G)}{qN_D}}\right]. \tag{29}$$

Equating this to Eq. 11, a transcendental equation for $V_{D\mathrm{sat}}$ is obtained

$$\mathscr{E}_cL = \frac{V_{D\mathrm{sat}} - [2/(3\sqrt{\psi_P})][(\psi_{bi} + V_{D\mathrm{sat}} - V_G)^{3/2} - (\psi_{bi} - V_G)^{3/2}]}{1 - \sqrt{(\psi_{bi} + V_{D\mathrm{sat}} - V_G)/\psi_P}}. \tag{30}$$

Visual examination of this equation indicates that current saturates as V_D approaches $\mathscr{E}_cL$, or $V_D/L \approx \mathscr{E}_c$. Once $V_{D\mathrm{sat}}$ is known, $I_{D\mathrm{sat}}$ can be calculated from Eq. 11 of the constant-mobility model. One also finds that since $V_{D\mathrm{sat}}$ is lower than the value from the constant-mobility model, current saturation occurs before the channel is pinched off.

Field-Dependent Mobility: Empirical Formula. We next derive the current equation based on the empirical $v(\mathscr{E})$ formula given in the form of Eq. 28. Substituting v into Eq. 9 and integrating from $x = 0$ to L, we obtain

$$\int_0^L I_D\left(1 + \frac{\mathscr{E}_x}{\mathscr{E}_c}\right)dx = \int_0^L ZQ_n\mu\mathscr{E}_x dx. \tag{31}$$

Notice that the right-hand side is similar to the constant-mobility model in Eq. 10 which results in Eq. 11. The left-hand side gives a value of $I_D(L + V_D/\mathscr{E}_c)$. After the integration Eq. 31 yields

$$I_D = \frac{G_i}{1 + (V_D/\mathscr{E}_cL)}\left\{V_D - \frac{2}{3\sqrt{\psi_P}}[(\psi_{bi} + V_D - V_G)^{3/2} - (\psi_{bi} - V_G)^{3/2}]\right\}. \tag{32}$$

Comparing this to Eq. 11, this new result gives a current that is reduced by a factor of $(1 + V_D/\mathscr{E}_cL)$ from that of the constant-mobility model. In order to obtain $V_{D\mathrm{sat}}$, we seek the current peak from Eq. 32 by setting $dI_D/dV_D = 0$. This yields a transcendental equation for $V_{D\mathrm{sat}}$ as

$$\mathscr{E}_cL = \sqrt{\frac{\psi_{bi} + V_{D\mathrm{sat}} - V_G}{\psi_P}}(\mathscr{E}_cL + V_{D\mathrm{sat}}) - \frac{2}{3\sqrt{\psi_P}}[(\psi_{bi} + V_{D\mathrm{sat}} - V_G)^{3/2} - (\psi_{bi} - V_G)^{3/2}]. \tag{33}$$

Solutions of $V_{D\mathrm{sat}}$ from the above equation have been calculated and plotted in Fig. 5, for various values of $\mathscr{E}_cL$. The top curve ($\mathscr{E}_cL = \infty$) becomes the limit of the constant-mobility model. Note that with decreasing $\mathscr{E}_cL$, the saturation of drain current is reached at smaller values of drain voltage.

To obtain the saturation drain current, the solutions for $V_{D\mathrm{sat}}$ can be used in Eq. 32, which is done by substituting some terms from Eq. 33 into Eq. 32:

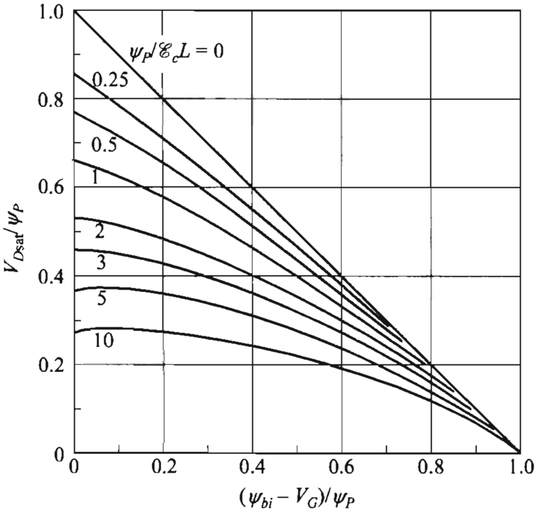


Fig. 5 Solution of V_{Dsat} from Eq. 33, for various values of $\psi_P/\mathscr{E}_cL$. (After Ref. 6.)

$$I_{Dsat} = G_i\mathscr{E}_cL\left(1 - \sqrt{\frac{\psi_{bi} + V_{Dsat} - V_G}{\psi_P}}\right) = G_i\mathscr{E}_cL(1 - u_{dm}), \tag{34}$$

where u_{dm} is the value of u_d evaluated at V_{Dsat}. The transconductance in the saturation region can then be obtained from taking derivatives of both Eqs. 33 and 34 (readers are reminded that V_{Dsat} is also a function of V_G):

$$g_m = \frac{dI_{Dsat}}{dV_G} = \frac{G_i}{\sqrt{\psi_P}}\left(\frac{\sqrt{\psi_{bi} + V_{Dsat} - V_G} - \sqrt{\psi_{bi} - V_G}}{1 + (V_D/\mathscr{E}_cL)}\right) = \frac{G_i(u_{dm} - u_s)}{1 + (\psi_P/\mathscr{E}_cL)(u_{dm}^2 - u_s^2)}. \tag{35}$$

This expression reduces to Eq. 27 of the constant-mobility model for $\mathscr{E}_cL = \infty$ and $u_{dm} = 1$.

Having gone through the three models of υ-$\mathscr{E}$ relationships, it is informative to compare their results on the I-V characteristics. Here we use an example of one I-V curve for a fixed V_G (= 0), and the other parameters are; $\psi_P = 4$ V, $\psi_{bi} = 1$ V, and $\mathscr{E}_cL = 2$ V. The results are shown in Fig. 6. The values of the V_{Dsat} for the constant mobility, two-piece liner approximation, and the empirical formula are 3 V, 1.3 V, and 1.9 V respectively. Note that the curve for the two-piece-linear model lies on the constant-mobility curve until V_{Dsat}. The lowest current of the three curves corresponds to the empirical formula of Eq. 29 since at any field, the velocity is the lowest among the three models as shown in Fig. 4.

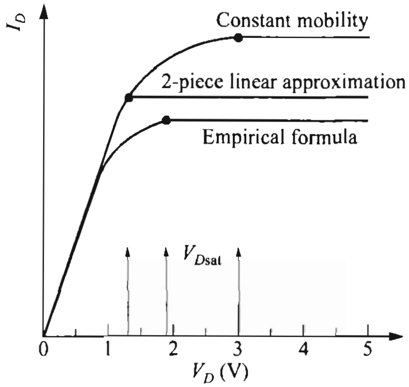


Fig. 6 I-V curves for a fixed V_G (= 0) for three models of υ-$\mathscr{E}$ relationship.

Velocity Saturation. One limiting case is the saturation-velocity model[7] which is expected to be valid in the limit of very short gates where $L \ll V_D/\mathscr{E}_c$. In this assumption, the carriers travel with υ_s in the whole region under the gate, and are totally independent of the low-field mobility. Starting from Eq. 9, the saturation current is simply given by

$$\begin{aligned} I_{\text{Dsat}} &= ZQ_n\upsilon_s \\ &= Zq(a - W_{Ds})N_D\upsilon_s . \end{aligned} \tag{36}$$

The maximum current for the devices is thus $ZqaN_D\upsilon_s$ which is reduced from that of the constant-mobility model given by $G_i\psi_P/3$. The choice of depletion width at the source W_{Ds} rather than at the drain is apparent as we discuss details of the carrier-density and velocity profiles in the next section (Dipole-Layer Formation). This equation shows an interesting feature that the saturation current is now totally independent of the channel length. The transconductance is given by

$$g_m = \frac{dI_{\text{Dsat}}}{dV_G} = -ZqN_D\upsilon_s\frac{dW_{Ds}}{dV_G} = \frac{Z\varepsilon_s\upsilon_s}{W_{Ds}} . \tag{37}$$

Since ε_s/W_{Ds} is the gate-source capacitance C_{GS}, this equation reduces to the familiar FET equation

$$g_m = ZC_{GS}\upsilon_s . \tag{38}$$

This equation also has the interesting feature that g_m is constant and totally independent of gate bias as well as channel length. Output characteristics of constant mobility and velocity saturation are compared in Fig. 7. Note that the saturation current and saturation voltage are lower under velocity saturation, but the linear regions remain similar. The constant g_m under velocity saturation is indicated by the equal spacing of the I-V curves under different V_G. As shown, this velocity-saturation limit provides very simple derivations and results, thus giving a good insight of the short-channel limit. In fact, the simple formulae can fit quite well to the state-of-the-art short-channel devices.

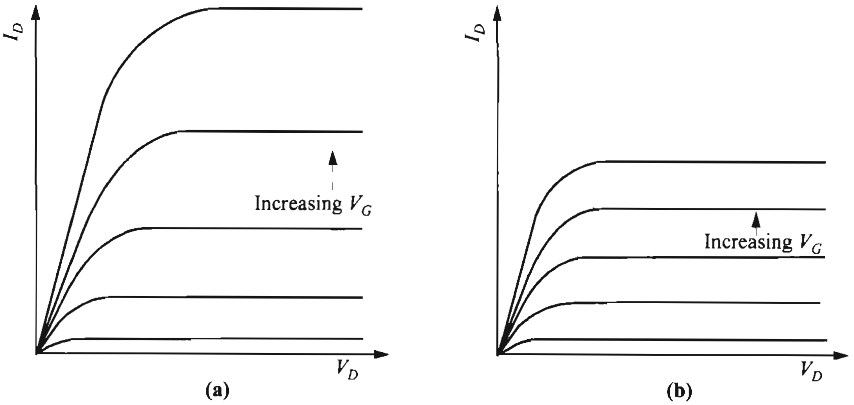


Fig. 7 A qualitative comparison of I-V curves under models of (a) constant mobility and (b) velocity saturation.

Even though velocity saturation sets a limit on the maximum carrier speed in a field-effect transistor, there are two special effects that enable higher speed at part of the channel where the local field is high. The first is related to the material properties, such as in GaAs and InP, which display a transferred-electron effect. According to the υ-$\mathscr{E}$ relationship shown in Fig. 20a (p. 38) in Chapter 1, at moderately high fields, the drift velocity is actually higher than the saturation velocity. To include this negative-resistance effect in modeling the I-V characteristics analytically would be very difficult. The second effect is present in ultra-short devices when their channel lengths are comparable to or smaller than the mean free path of scattering. Readers are referred to the discussion of ballistic effect in Chapter 1 (p. 37). For very short gates, the electrons may not have enough time or distance to reach equilibrium transport in the high-field region of the channel.[8] In such cases, the electrons enter the high-field region and are accelerated to a higher velocity before relaxing to the equilibrium value. Carriers can thus overshoot to more than twice the steady-state velocity and then relax to the equilibrium condition after traveling a certain distance. The overshoot will shorten the electron transit time. This overshoot is expected to improve high-frequency response, especially for the GaAs FET. This phenomenon is related indirectly to low-field mobility since they are both determined by scattering. A material with higher mobility can have more ballistic effect for the same channel length.

Dipole-Layer Formation. An interesting phenomenon occurs that is associated with velocity saturation and when biased beyond V_{Dsat}. This stems from the fact that as drain bias increases beyond V_{Dsat}, the depletion layer continues to grow and meanwhile the net channel opening is reduced. In order to maintain the same saturation current, the carriers concentration in the narrower channel has to rise above the doping level in order to maintain the same current since velocity is fixed at υ_s. This is explained in more details as follows.

Below the saturation drain bias $V_{D\text{sat}}$, the potential along the channel increases from zero at the source to V_D at the drain. Thus, the gate contact becomes increasingly reverse-biased with respect to the channel, and the depletion width becomes wider as we proceed from source to drain. The resulting decrease in channel opening b must be compensated by an increase in electric field and electron velocity to maintain a constant current throughout the channel. As V_D is increased further to $V_{D\text{sat}}$, the electrons reach the saturation velocity at the drain end of the gate (Fig. 8a). The channel is constricted to the smallest cross section b_1 under the gate. The electric field reaches the critical value $\mathscr{E}_c$ at this point, and I_D starts to saturate. The electron density $n(x)$, however, remains equal to the doping density N_D as long as the field does not exceed the critical value $\mathscr{E}_c$.

The condition of $V_D > V_{D\text{sat}}$ is displayed in Fig. 8b. The saturation current is given by

$$I_{D\text{sat}} = Zqv_s n(x) b(x). \tag{39}$$

If the drain voltage is increased beyond $V_{D\text{sat}}$, the depletion region widens toward the drain. The point x_1, where the electrons reach the saturation velocity and the channel width is b_1, moves toward the source. Note that there are three locations of interest;

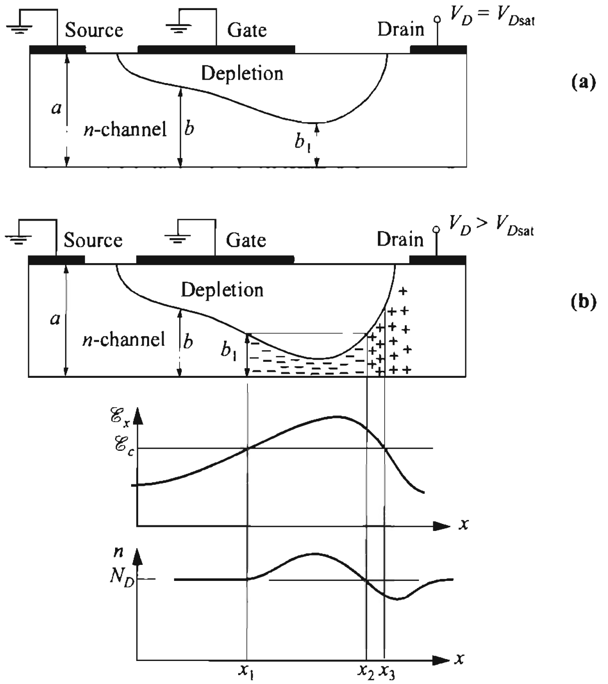


Fig. 8 (a) Schematic cross-section showing condition at $V_D = V_{D\text{sat}}$ and under velocity saturation. (b) Dipole-layer formation when operated with $V_D > V_{D\text{sat}}$, showing electric-field and carrier-concentration profiles through the quasi-neutral channel. (After Ref. 9.)

x_1 and x_2 are locations where the channel opening is b_1, and x_1 and x_3 are locations where $\mathscr{E} = \mathscr{E}_c$. This also means that in the region x_1 to x_2 the channel is narrower than b_1, and in the region x_1 to x_3, carriers travel with v_s. Since the velocity is saturated, in the region x_1 to x_2 the change in channel width must be compensated by a change in carrier density to maintain constant current. According to Eq. 39, an electron accumulation layer ($n > N_D$) must form in this region, where the channel opening is smaller than b_1. At x_2 the channel opening is again b_1, and the negative space charge changes to a positive space charge ($n < N_D$) to preserve constant current. Between x_2 and x_3 the electron velocity remains saturated but the channel width is larger than b_1. So by virtue of the same Eq. 39, carrier concentration is lower than N_D in order to keep the saturation current constant. Therefore, the drain voltage applied in excess of V_{Dsat} forms a dipole layer in a channel that extends beyond the drain end of the gate.

Breakdown. For drain voltages beyond $V_{D\text{sat}}$, the drain current is assumed to remain essentially the same as the saturation current. As the drain voltage increases further, breakdown occurs where the current rises sharply with the drain bias. This breakdown occurs at the gate edge toward the drain side where the field is the highest. Analysis of the breakdown condition in an FET is inherently more complicated than in a bipolar transistor because it is a two-dimensional situation as opposed to one-dimensional.

The fundamental mechanism responsible for breakdown is impact ionization. Since impact ionization is a strong function of the electric field, the maximum field is often regarded as the first-order criterion for breakdown. Using a simple one-dimensional analysis in the x-direction, and treating the gate-drain structure as a reverse-biased diode, the drain breakdown voltage V_{DB} is similar to that of the gate-junction breakdown and is linearly dependent on the relative voltage of the drain to the gate;

$$V_{DB} = V_B - V_G \tag{40}$$

where V_B the breakdown voltage of the gate diode and is a function of channel doping level, among other factors. The general breakdown behavior of Eq. 40 is shown in Fig. 9a. It is shown that for higher V_G, the drain breakdown voltage becomes higher by the same amount. Such characteristics general hold true for silicon JFETs. But for MESFETs on GaAs, the breakdown mechanisms are much more complicated. They generally have lower breakdown values and the dependence of V_G no longer follows Eq. 40 but has a opposite trend, as shown in Fig. 9b. These additional effects will be discussed next.

Unlike a MOSFET where the heavily doped source and drain overlap the gate at the gate edges, the JFETs and MESFETs have a gap between the gate and the source/drain contacts (or heavily doped regions under the contacts). For breakdown consideration, this gate-drain distance L_{GD} is critical. In this gap the doping level is the same as the channel. If surface traps are present in this gate-drain spacing, they can deplete part of the channel doping and affect the field distribution. In certain cases they can improve the breakdown voltage. Two-dimensional simulations in Fig. 10 displays the field distribution as a function of surface potential created by surface traps. Without surface traps (i.e., $\psi_s = 0$), the field is highest at the gate edge where

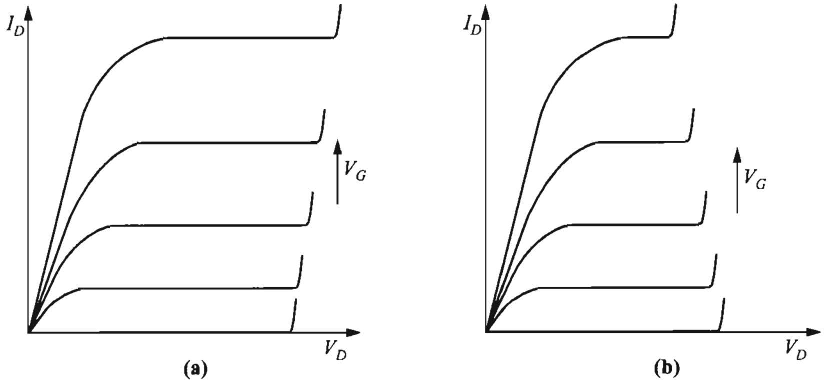


Fig. 9 Experimental data showing drain breakdown voltage (a) increases with V_G in Si JFETs but (b) decreases with V_G in GaAs MESFETs.

breakdown occurs. In this particular example, with a surface potential of 0.65 V, the field at the gate edge is reduced, thereby increasing the breakdown voltage. Using a one-dimensional analysis, the field at the gate edge can be shown to be[10]

$$\mathscr{E}(L) = \frac{qN_D}{\varepsilon_s}\sqrt{\frac{2\varepsilon_s}{qN_D}(\psi_{bi}+V_D-V_G)-\frac{N'_{st}}{N_D}L^2_{GD}} \tag{41}$$

where N'_{st} is the surface-trap density. (This equation implies that L_{GD} is larger than the 1-D depletion width so that with $N'_{st} = 0$, $\mathscr{E}(L)$ and V_{DB} are independent of L_{GD}.) For an increased surface potential of 1.0 V, the field at the drain contact is increased, since the area under the curve is the total applied voltage and it has to be conserved. If the field at the drain contact is increased to a critical value, breakdown can occur there, thereby lowering the breakdown voltage again. Since GaAs lacks a common

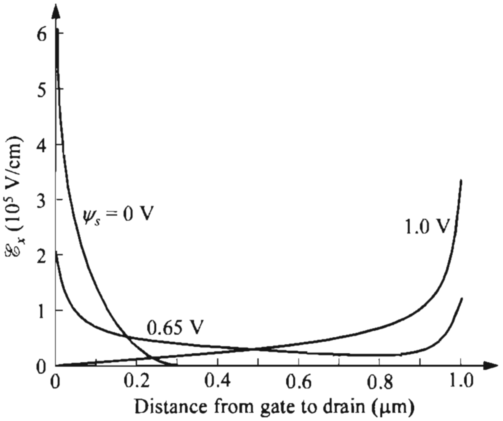


Fig. 10 Electric-field distribution in gate-drain spacing as a function of surface potential ψ_s due to surface traps. $V_D = 4$ V. $V_G = 0$. (After Ref. 10.)

passivation layer such as SiO_2 for silicon, the breakdown in GaAs MESFETs are less controllable and have different breakdown behavior compared to Si JFETs.

One factor for a reduced breakdown voltage in MESFETs is due to tunneling current associated with the Schottky-barrier gate contact.[11] At high fields, this tunneling current is from thermionic-field emission which has a temperature dependence. The gate current can initiate avalanche multiplication and induces lower drain breakdown voltage. With higher channel current, the internal node is at a higher temperature which triggers an earlier gate-current-initiated avalanche breakdown. This can be responsible for lower V_{DB} at higher V_G in Fig. 9b. Another factor is that GaAs MESFETs usually have higher current and transconductance than Si devices due to higher mobility. The higher channel current can initiate avalanche at a lower voltage, or produce the temperature effects which trigger earlier breakdown as discussed.

The breakdown voltage can be improved by extending the region between the gate and the drain. Furthermore, to maximize its function, the field distribution should be made as uniform as possible. One technique is to introduce a doping gradient in the lateral direction. Another, called RESURF (reduced surface field),[12] is to have a p-layer underneath such that at high drain bias, this n-layer is fully depleted.

7.2.2 Arbitrary Doping and Enhancement Mode

Arbitrary Doping Profile. For an arbitrary doping profile in the channel region,[13] the net potential variation inside the depletion width is related to the doping as given by Eq. 40 of Chapter 2,

$$\psi_{bi} - V_G = \frac{q}{\varepsilon_s}\int_0^{W_D} y N_D(y)dy. \tag{42}$$

The maximum value for the upper limit of the integral occurs at $W_D = a$, and the corresponding quantity is the pinch-off potential as defined previously, given by

$$\psi_P = \frac{q}{\varepsilon_s}\int_0^{a} y N_D(y)dy. \tag{43}$$

We next consider the current-voltage characteristics and the transconductance. We shall define an integral form of the total charge density up to the position y_1 as

$$Q(y_1) \equiv q\int_0^{y_1} N_D(y)dy, \tag{44}$$

which will be used to simplify the equations that follow. The drain current based on Eq. 9 would have to be modified to be

$$\begin{aligned} I_D &= Zqv\int_{W_D}^{a} N_D(y)dy \\ &= Zv[Q(a) - Q(W_D)]. \end{aligned} \tag{45}$$

Bear in mind that both v and W_D vary with x along the channel under a drain bias. Integrating both sides of this equation from $x = 0$ to L gives

$$\int_0^L I_D dx = I_D L = Z\int_0^L v[Q(a) - Q(W_D)]dx . \tag{46a}$$

or

$$I_D = \frac{Z}{L}\int_0^L v[Q(a) - Q(W_D)]dx . \tag{46b}$$

Equation 46b is the basic equation for calculating the drain current.

In the linear region, the drift velocity is always in the constant-mobility regime due to the small field or drain bias. Substituting $v = \mu\mathscr{E} = \mu d\Delta\psi_i/dx$, we obtain

$$I_{D\text{lin}} = \frac{Z}{L}\int_0^L \mu\frac{d\Delta\psi_i}{dx}[Q(a) - Q(W_D)]dx = \frac{Z\mu}{L}\int_0^{V_D} [Q(a) - Q(W_D)]d\Delta\psi_i$$
$$\approx \frac{Z\mu}{L}[Q(a) - Q(W_{Ds})]V_D \ . \tag{47}$$

In the saturation region, we first consider the case where saturation is caused by pinch-off ($W_{Dd} = a$) as opposed to velocity saturation. Starting again with Eq. 46b and changing the variable to W_D with Eq. 21, the drain current is

$$I_{D\text{sat}} = \frac{Z\mu}{L}\int_{W_{Ds}}^{a} [Q(a) - Q(W_D)]\frac{d\Delta\psi_i}{dW_D}dW_D$$
$$= \frac{Zq\mu}{\varepsilon_s L}\int_{W_{Ds}}^{a} [Q(a) - Q(W_D)]\, W_D N_D dW_D . \tag{48}$$

Using the relationship similar to Eq. 21,

$$\frac{dW_D}{dV_G} = \frac{-\varepsilon_s}{qW_D N_D} , \tag{49}$$

the transconductance is given by differentiating Eq. 48

$$g_m = \frac{dI_{D\text{sat}}}{dV_G} = \frac{dI_{D\text{sat}}}{dW_D} \times \frac{dW_D}{dV_G} = \frac{-Zq\mu}{\varepsilon_s L}[Q(a) - Q(W_{Ds})]\, W_D N_D \times \frac{dW_D}{dV_G}$$
$$= \frac{Z\mu}{L}[Q(a) - Q(W_{Ds})] \ , \tag{50}$$

which shows that g_m is equal to the conductance of the rectangular section of the semiconductor extending from $y = W_{Ds}$ to a.

For short-channel devices where velocity saturation determines current saturation, the drain current is simply given by

$$I_{D\text{sat}} = Zqv_s\int_{W_{Ds}}^{a} N_D(y)dy = Zv_s[Q(a) - Q(W_{Ds})]. \tag{51}$$

To obtain the transconductance, Eq. 51 implies

$$\frac{dI_{D\text{sat}}}{dW_{Ds}} = -Zqv_sN_D(W_{Ds}). \tag{52}$$

The transconductance is given by

$$\begin{aligned} g_m &= \frac{dI_{D\text{sat}}}{dW_{Ds}} \times \frac{dW_{Ds}}{dV_G} = -Zqv_sN_D \times \frac{-\varepsilon_s}{qW_{Ds}N_D} \\ &= \frac{Zv_s\varepsilon_s}{W_{Ds}}, \end{aligned} \tag{53}$$

which is identical to Eq. 37.

In real applications, it is often preferable to have good linearity, i.e., constant g_m, meaning $I_{D\text{sat}}$ changes linearly with V_G. Linearity of the transfer characteristics is approached by those profiles in which the depletion depth $W_D(V_G)$ changes very little as a function of the gate voltage. The transfer characteristics for various doping profiles are shown in Fig. 11. Note that both types of nonuniform dopings achieve linearity as the appropriate variable parameter is taken to its limit, which has a delta doping at $x = a$. The results shown are quite different from the constant-mobility case, in which the doping profile has little effect on the transfer characteristics. Although Eq. 53 implies a reduction of g_m for lower gate voltage, the important quantity g_m/C_{GS} remains unaffected, where C_{GS} is the gate-source capacitance. This is because $C_{GS} = \varepsilon_s/W_D$, and Eq. 53 gives

$$\frac{g_m}{C_{GS}} = Zv_s = \text{constant}. \tag{54}$$

Experimental results have confirmed that FETs with graded channel doping[14] or step dopings[15] have improved linearity.

Enhancement-Mode Devices. Buried-channel FETs are usually normally-on devices. The basic current-voltage characteristics of normally-on and normally-off devices are similar, except for the value of the threshold voltage. Figure 12 compares these two modes of operation. The main difference is the shift of threshold voltage along the V_G-axis. The normally-off device has no current conduction at $V_G = 0$, and when $V_G > V_T$ the current starts to flow.

For high-speed low-power applications, the normally-off (or enhancement-mode) device is very attractive. A normally-off device is one that does not have a conductive channel at $V_G = 0$; that is, the built-in potential ψ_{bi} of the gate junction is sufficient to totally deplete the channel region. Mathematically, normally-off device has a positive V_T, implying from Eq. 15

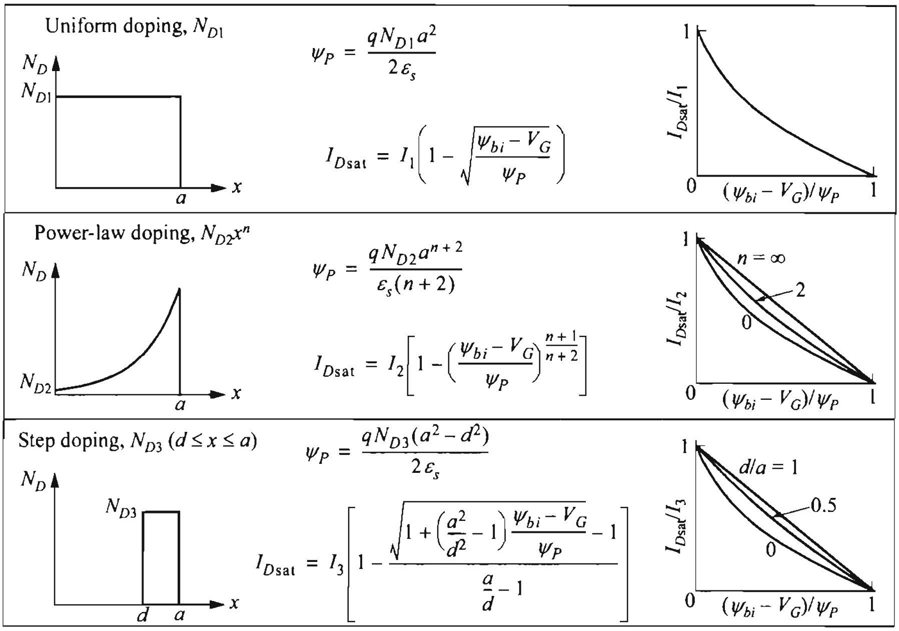

Uniform doping, N_{D1}
N_D
N_{D1}
a
x
$\psi_P = \frac{qN_{D1}a^2}{2\varepsilon_s}$
$I_{Dsat} = I_1\left(1 - \sqrt{\frac{\psi_{bi} - V_G}{\psi_P}}\right)$
I_{Dsat}/I_1
1
0
$(\psi_{bi} - V_G)/\psi_P$
1
Power-law doping, $N_{D2}x^n$
N_D
N_{D2}
a
x
$\psi_P = \frac{qN_{D2}a^{n+2}}{\varepsilon_s(n+2)}$
$I_{Dsat} = I_2\left[1 - \left(\frac{\psi_{bi} - V_G}{\psi_P}\right)^{\frac{n+1}{n+2}}\right]$
I_{Dsat}/I_2
1
$n = \infty$
2
0
0
$(\psi_{bi} - V_G)/\psi_P$
1
Step doping, N_{D3} $(d \le x \le a)$
N_D
N_{D3}
d
a
x
$\psi_P = \frac{qN_{D3}(a^2 - d^2)}{2\varepsilon_s}$
$I_{Dsat} = I_3\left[1 - \frac{\sqrt{1 + \left(\frac{a^2}{d^2} - 1\right)\frac{\psi_{bi} - V_G}{\psi_P}} - 1}{\frac{a}{d} - 1}\right]$
I_{Dsat}/I_3
1
$d/a = 1$
0.5
0
0
$(\psi_{bi} - V_G)/\psi_P$
1

Fig. 11 I_{Dsat} expressions and transfer characteristics for various doping profiles. Velocity-saturation model is assumed. (After Ref. 7.)

$$\begin{aligned} \psi_{bi} &> \psi_P \\ &> \frac{qN_Da^2}{2\varepsilon_s} . \end{aligned} \tag{55}$$

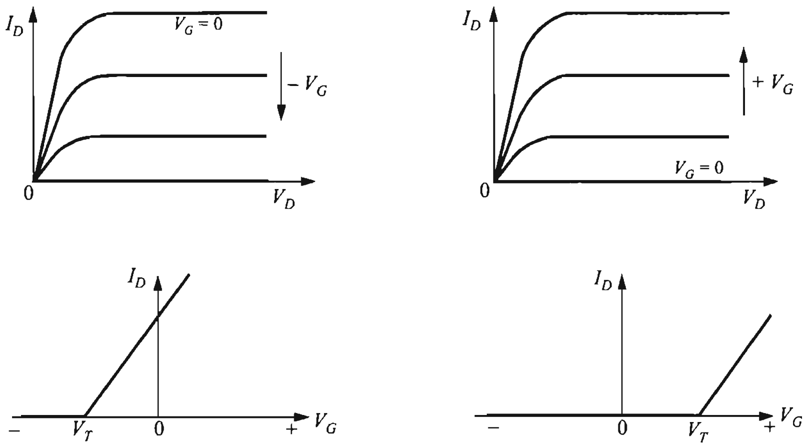
Since the built-in potential ψ_{bi} has a limit comparable to the energy gap, it imposes a limit on the channel doping and channel width, both of which affect the maximum current the device can provide. For a uniformly doped channel that is saturation-velocity limited, the maximum current is given by

$$I_D < ZqN_Dav_s . \tag{56}$$

This current limit would be obtained if the applied gate bias were equal to the built-in potential, an impractical biasing condition which will cause excessive gate current.

7.2.3 Microwave Performance

Small-Signal Equivalent Circuit. Field-effect transistors, especially GaAs MESFETs, are useful for low-noise amplification, high-efficiency power generation, and high-speed logic applications. We shall first consider the small-signal equivalent circuit of a MESFET or JFET. A small-signal lumped-element circuit for operation in the saturation region in common-source configuration is shown in Fig. 13. In the



(a) **(b)**

Fig. 12 Comparison of I-V characteristics for (a) normally-on (depletion-mode) FET and (b) normally-off (enhancement-mode) FET.

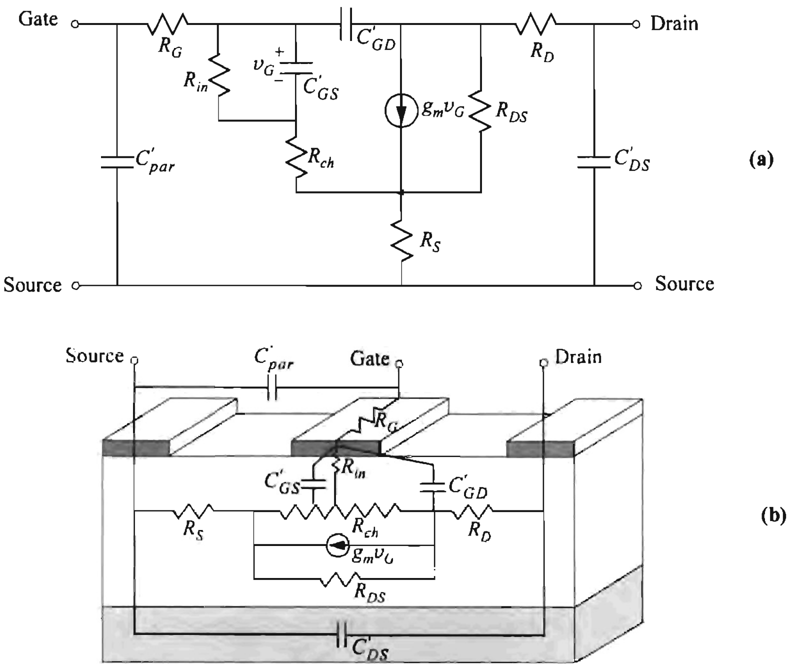

Gate
R_G
R_{in}
v_G
+
−
C'_{GD}
C'_{GS}
$g_m v_G$
R_{DS}
R_D
Drain
C'_{par}
R_{ch}
C'_{DS}
(a)
R_S
Source
Source
Source
C'_{par}
Gate
Drain
R_G
R_{in}
C'_{GS}
C'_{GD}
R_S
R_{ch}
R_D
$g_m v_G$
R_{DS}
C'_{DS}
(b)

Fig. 13 (a) Small-signal equivalent circuit of MESFET and JFET. v_G is small-signal V_G. Capacitance symbol with prime designates total capacitance in farad as opposed to capacitance per unit area. (b) Origin of circuit elements is related to physical structures.

intrinsic FET, the elements $C'_{GS} + C'_{GD}$ are the total gate-channel capacitance ($= C'_G$); R_{ch} is the channel resistance; R_{DS} is the output resistance which reflects the nonsaturating drain current with drain bias. The extrinsic (parasitic) elements include the source and drain series resistances R_S and R_D, the gate resistance R_G, the parasitic input capacitance C'_{par} and output (drain-source) capacitance C'_{DS}.

The leakage current in the gate-to-channel junction can be expressed as

$$I_G = I_0\left[\exp\left(\frac{qV_G}{nkT}\right) - 1\right] \tag{57}$$

where n is the diode ideality factor ($1 < n < 2$) and I_0 is the saturation current. The input resistance is given by

$$R_{in} \equiv \left(\frac{dI_G}{dV_G}\right)^{-1} = \frac{nkT}{q(I_0 + I_G)}. \tag{58}$$

As I_G approaches zero, the input resistance at room temperature is about 250 MΩ for $I_0 = 10^{-10}$ A. It becomes even higher for negative gate bias (negative I_G). The FET obviously has a very high input resistance, even though it is not as ideal as in a MOSFET which has an insulating gate.

The source and drain series resistances, which cannot be modulated by the gate voltage, will introduce an IR drop between the gate and the source and drain contacts. These IR drops will reduce the drain conductance as well as the transconductance. The internal effective voltages V_D and V_G should then be replaced by $[V_D - I_D(R_S+R_D)]$ and $(V_G - I_D R_S)$, respectively. In the linear region, the resistances R_S and R_D are in series, adding to the total measured drain-source resistance $(R_S+R_D+R_{ch})$. In the saturation region, the drain resistance R_D will cause an increase of the drain voltage at which current saturation occurs. Beyond that voltage $V_D > V_{Dsat}$, the magnitude of V_D has no effect on the drain current. By the same token, the measured transconductance in the saturation region is affected only by the source resistance. Thus R_D has no further effect on g_m, and the measured extrinsic transconductance is equal to

$$g_{mx} = \frac{g_m}{1 + R_S g_m}. \tag{59}$$

Cutoff Frequency. For a measure of the high-speed capability, the cutoff frequency f_T and the maximum frequency of oscillation $f_{\max}$ are commonly used. The f_T is defined as the frequency of unity gain, at which the small-signal input gate current is equal to the drain current of the intrinsic FET. The $f_{\max}$ is the maximum frequency at which the device can provide power gain. The f_T is a more appropriate figure-of-merit for digital circuits where speed is the primary concern, and $f_{\max}$ is more relevant for analog-circuit applications.

Based on unity gain, one can use the derivation discussed in Section 5.3.1 (p. 263) for an expression

$$f_T = \frac{g_m}{2\pi C'_{in}} = \frac{g_m}{2\pi(C'_G + C'_{par})}. \tag{60}$$

Here C'_{in} is the total input capacitance (Fig. 13), and C'_G is the sum of C'_{GS} and C'_{GD}. For an ideal case of zero input parasitics ($C'_{par} = 0$), we obtain

$$f_T = \frac{g_m}{2\pi C'_G} = \frac{v}{2\pi L}. \tag{61}$$

This equation has the physical meaning that f_T is related to the ratio L/v which happens to be the transit time for a carrier to travel from source to drain. The drift velocity v is equal to the saturation velocity v_s for short channels, and for a 1-μm gate length, the transit time is of the order of 10 ps (10^{-11} s). In practice, the parasitic input capacitance C'_{par} is a fraction of C'_G, so f_T is slightly below its theoretical maximum value.

Equation 60 is an approximation ignoring some of the parasitics. A more-complete equation accounting for source and drain resistances and the gate-drain capacitance is given by

$$f_T = \frac{g_m}{2\pi\left[C'_G\left(1 + \frac{R_D + R_S}{R_{DS}}\right) + C'_{GD}g_m(R_D + R_S) + C'_{par}\right]}. \tag{62}$$

Note that g_m is the intrinsic value but not g_{mx} as in Eq. 59.

The speed limitations of FETs are also dependent on device geometry and material properties. In the device geometry, the most-important parameter is the gate length L. Decreasing L will decrease the total gate capacitance $[C'_G \propto (Z \times L)]$ and increase the transconductance (before velocity saturation); consequently, f_T improves. As for the carrier transport, since the internal field varies in magnitude along the channel, drift velocities in all field strength are critical. These include the low-field mobility, saturation velocity in high field, and for some materials, peak velocity in medium field in the presence of the transferred-electron effect. In Si and GaAs, electrons have a higher low-field mobility than holes have. Therefore only *n*-channel FETs are used in microwave applications. The low-field mobility in GaAs is about five times higher than that of silicon, therefore the frequency f_T is expected to be higher in GaAs. For the same gate length, InP is expected to have even higher f_T than GaAs because of its higher peak velocity. In any case, for these materials, FETs with gate lengths 0.5 μm or less will have f_T in excess of 30 GHz.

Maximum Frequency of Oscillation. The $f_{\max}$ is defined as the frequency at which the unilateral gain is unity. The unilateral gain U goes down as square of frequency,

$$U \approx \left(\frac{f_{\max}}{f}\right)^2, \tag{63}$$

with

$$f_{\max} = \frac{f_T}{2\sqrt{r_1 + f_T \tau_3}}. \tag{64}$$

r_1 is the input-to-output resistance ratio,

$$r_1 \equiv \frac{R_G + R_{ch} + R_S}{R_{DS}}, \tag{65}$$

and the channel resistance R_{ch} is given by[16]

$$R_{ch} = \frac{1}{g_m} \frac{(3\alpha^3 + 15\alpha^2 + 10\alpha + 2)(1 - \alpha)}{10(1 + \alpha)(1 + 2\alpha)^2}. \tag{66}$$

α is a measure of the drain bias with respect to V_{Dsat},

$$\alpha = 1 - \frac{V_D}{V_{Dsat}} \qquad (V_D \leq V_{Dsat}). \tag{67}$$

So for the saturation region, $\alpha = 0$ and $R_{ch} = 1/5g_m$. The τ_3 is a time constant

$$\tau_3 \equiv 2\pi R_G C'_{GD}. \tag{68}$$

For small r_1, Eq. 64 reduces to the more-familiar form

$$f_{\max} \approx \sqrt{\frac{f_T}{8\pi R_G C'_{GD}}}. \tag{69}$$

The unilateral gain will decrease at 6 dB/octave as the frequency increases. At $f_{\max}$, unity power gain is reached. To maximize $f_{\max}$, the frequency f_T and the resistance ratio R_{ch}/R_{DS} must be optimized in the intrinsic FET. In addition, the extrinsic resistances R_G and R_S and the feedback capacitance C'_{GD} must also be minimized.

Power-Frequency Limitations. For power applications, both high voltage and high current are required. These, however, demand device designs that are in conflict with one another, and in addition, they also compromise the speed, so a trade-off has to be considered. For high current, the total channel dose ($N_D \times a$) has to be high. To maintain high breakdown voltage, N_D cannot be too high and L cannot be too small. For a high f_T, L has to be minimized and as a consequence, N_D has to increase. The last constraint comes about because of the following.

For a gate electrode to have adequate control of the current transport across the channel, the gate length must be somewhat larger than the channel depth,[17] that is

$$\frac{L}{a} \geq \pi. \tag{70}$$

So to reduce L, the channel depth a has to be reduced at the same time, which implies a higher doping level to maintain a reasonable current. Because of this, some scaling rules have been proposed. These include constant LN_D scaling, constant $L^{1/2}N_D$ scaling.[18] and constant L^2N_D scaling.[19] In practical Si and GaAs MESFETs, the highest doping level is about 5×10^{17} cm^{-3} because of breakdown phenomena. Using the simple velocity-saturation estimate of $I_{Dsat}/Z = qN_D a v_s$, and v_s of 1×10^7 cm/s, to maintain a current of 3 A/cm, this doping level limits the minimum gate length to about 0.1 μm, with a corresponding maximum f_T of the order of 100 GHz.

In high-power operation, the device temperature increases. This increase causes a reduction of the mobility[20] and saturation velocity, since the mobility varies as

$[T(\mathrm{K})]^{-2}$ and velocity as $[T(\mathrm{K})]^{-1}$. Therefore, the FET has negative temperature coefficient and will be thermally stable under high-power operation.

The state-of-the-art power-frequency performance of GaAs FETs is shown in Fig. 14. A higher frequency range can be reached with MODFETs, at the expense of lower power. With further miniaturization to submicron dimensions, and with improved designs and reduction of parasitics, FETs of higher powers operated at higher frequencies can be made. Also with semiconductor materials of higher bandgaps such as SiC and GaN, the curve can be shifted upward. For GaN devices, the curve can be higher by more than tenfold in power.[22]

Noise Behavior. The MESFET and JFET are low-noise devices, because only majority carriers participate in their operations, and these carriers transport through the channel in the bulk and free of surface or interface scattering. However, in practical devices, extrinsic resistances are unavoidable, and these parasitic resistances are mainly responsible for the noise behavior.

The equivalent circuit used for noise analysis is shown in Fig. 15. Noise sources i_{ng}, i_{nd}, e_{ng}, and e_{ns} represent the induced gate noise, induced drain noise, thermal noises of the gate resistance R_G and source resistance R_S, respectively. The e_s and Z_s are the signal source voltage and source impedance. The circuit within the dashed lines corresponds to the intrinsic FET. The noise figure is defined as the ratio of the total noise power to the noise power generated from the source impedance alone. So the noise figure depends also on the circuitry external to the device. There is an important parameter called the *minimum noise figure* which is obtained with both the source impedance and load impedance being optimized for noise performance. This minimum noise figure for a practical device has been obtained from the equivalent circuit to be:[24]

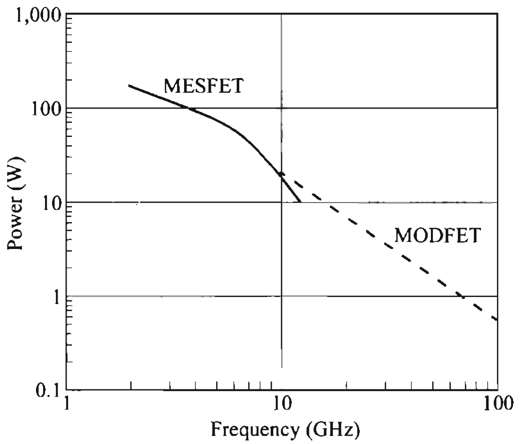


Fig. 14 State-of-the-art performance of power vs. frequency for GaAs FETs. Higher frequency range can be reached with MODFETs. (After Ref. 21.)

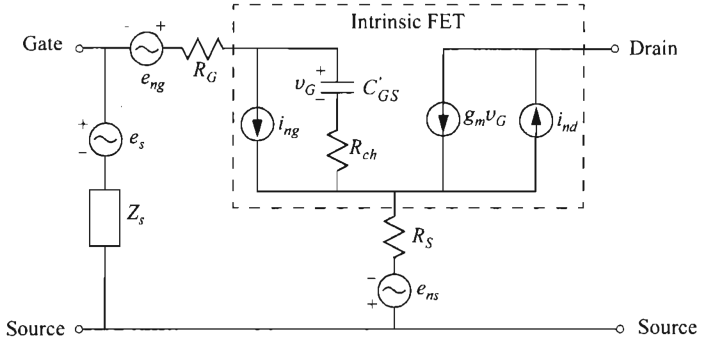


Fig. 15 Equivalent circuit of an FET for noise analysis. (After Ref. 23.)

$$F_{\min} \approx 1 + 2\pi C_1 f C'_{GS} \sqrt{\frac{R_G + R_S}{g_m}}, \tag{71}$$

where C_1 is a constant of value 2.5 s/F. Clearly for low-noise performance, one should minimize the parasitic gate resistance and source resistance. At a given frequency, the noise decreases with decreasing gate length ($C'_{GS} \propto L \times Z$). We should be reminded that R_G (see Fig. 13b) and g_m are proportional to the device width Z, while R_S is inversely proportional to Z. This leads to decreasing noise with decreasing channel width.

The graded-channel FET (Fig. 11) has been found to yield less noise (1 to 3 dB lower) than uniformly doped devices having the same geometry.[7] This difference in noise is related to g_m. The reduction of g_m (but not g_m/C'_{GS} for f_T) for a graded-channel FET gives superior noise performance.

7.2.4 Device Structures

The schematic diagrams of high-performance MESFETs are shown in Fig. 16. The MESFET structures fall into two major categories: the ion-implanted planar structure and the recessed-channel (or recessed-gate) structure. All devices have a semiinsulating (SI) substrate for compound semiconductors such as GaAs.

In the ion-implanted planar process (Fig. 16a), the active region is created by ion implantation to over-compensate the deep-level impurities in the SI-substrate. Naturally the active device is isolated vertically and horizontally by the semiinsulating material. To minimize the source and drain parasitic resistance, the deeper n^+-implantations should be as close to the gate as possible. This is done by various self-aligned processes. In a gate-priority self-aligned process, the gate is formed first, and the source/drain ion implantation is self-aligned to the gate. In this process, since ion implantation requires high-temperature anneal for activation, the gate has to made of materials that can withstand high-temperature processing. Examples are Ti-W alloy, WSi_2, and $TaSi_2$. The second approach is ohmic-priority where the source/drain

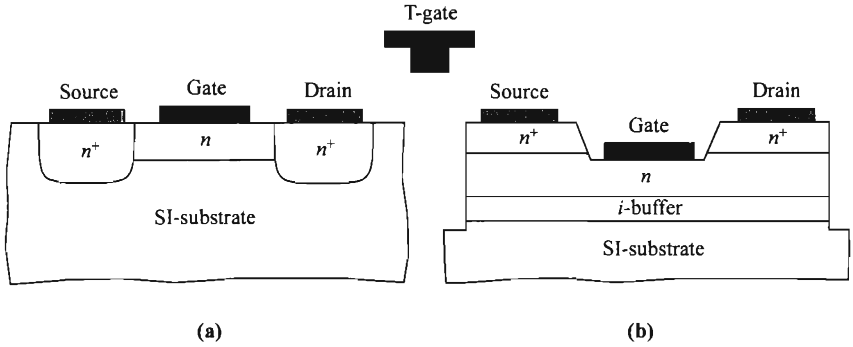


Fig. 16 Basic MESFET structures: (a) Ion-implanted planar structure. (b) Recessed-channel (or recessed-gate) structure. Inset shows a T-gate (or mushroom-gate) that can be used in both configurations.

implantation and anneal are done before the gate formation. Such process relaxes the previous requirement on the gate material.

In the recessed-channel process (Fig. 16b), the active layers are grown epitaxially over the SI-substrate. An intrinsic buffer layer is first grown and followed by an active channel layer. The buffer layer serves to eliminate defects duplicating from the semiinsulating substrate. Finally an epitaxial n^+-layer is grown over the active n-channel to reduce the source and drain contact resistance. This n^+-layer is selectively removed in the region between the source and the drain for gate formation. Sometimes, this etching process is monitored by measuring the current between source and drain for a more-precise control of the final channel current. One of the advantages of this recessed-channel structure is that the surface is further away from the n-channel layer so that surface effects such as transient response[25] and other reliability problems are minimized. One disadvantage of this scheme is the additional steps required for isolation which could be a mesa etching process (as shown) or an isolation implantation that converts the semiconductor into high-resistivity material.

For superior microwave performance, the gate can be made into the shape of a T-gate or mushroom-gate as shown in the inset of Fig. 16. The shorter dimension at the bottom of the gate is the electrical channel length and it serves to optimize f_T and g_m, while the wider top portion reduces the gate resistance for an improved $f_{\max}$ and noise figure.

The JFET structures are similar to those of the MESFET, with the additional step of a p-n junction formed underneath the gate contact by ion implantation. JFETs are more suitable for power applications and seldom used in state-of-the-art high-frequency applications. Part of the reason is that the channel length from a p-n junction is more different to control and miniaturize compared to a metal gate. One common inherent shortcoming of both the MESFET and JFET is that the heavily doped source and drain regions cannot overlap the gate as in the case of a MOSFET (see Fig. 6 on

p. 298). If they encroach the gate underneath, a short or leaky path would be formed between the gate and the source or drain. Because of this, the source and drain series resistance is higher than that in a MOSFET.

7.3 MODFET

The modulation-doped field-effect transistor (MODFET) is also known as the high-electron-mobility transistor (HEMT), two-dimensional electron-gas field-effect transistor (TEGFET), and selectively doped heterojunction transistor (SDHT). Sometimes it is simply referred to by the generic name HFET (heterojunction field-effect transistor). The unique feature of the MODFET is the heterostructure, in which the wide-energy-gap material is doped and carriers diffuse to the undoped narrow-bandgap layer at which heterointerface the channel is formed. The net result of this modulation doping is that channel carriers in the undoped heterointerface are spatially separated from the doped region and have high mobilities because there is no impurity scattering.

Carrier transport parallel to the layers of a superlattice was first considered by Esaki and Tsu in 1969.[26] The development of MBE and MOCVD technologies in the 1970s made heterostructures, quantum wells, and superlattices practical and more accessible. Dingle et al. first demonstrated enhanced mobility in the AlGaAs/GaAs modulation-doped superlattice in 1978.[27] Stormer et al. subsequently reported similar effect using a single AlGaAs/GaAs heterojunction in 1979.[28] These studies were made on two-terminal devices without the control gate. This effect was applied to the field-effect transistor by Mimura et al. in 1980,[29,30] and later by Delagebeaudeuf et al. in the same year.[31] Since then, the MODFET has been the subject of major research activities and has matured to commercial products as an alternative to MESFETs in high-speed circuits. For in-depth treatment of the MODFET, readers are referred to Refs. 32–35.

The main advantage of modulation doping is the superior mobility. This phenomenon is demonstrated in Fig. 17 which compares mobilities in the bulk, relevant for MESFETs and JFETs, to that in the modulation-doped channel. It is seen here that since in a MESFET or JFET the channel has to be doped to a reasonably high level ($> 10^{17}$ cm^{-3}), the modulation-doped channel has much higher mobilities at all temperatures. It is also interesting to compare the modulation-doped channel, which usually has an unintentional doping below 10^{14} cm^{-3}, to lowly doped bulk samples, since in this case their impurity concentrations are similar. The bulk mobility as a function of temperature shows a peak but drops at both high temperature and low temperature (see Section 1.5.1 on p. 28). The decrease of bulk mobility with increase of temperature is due to phonon scattering. At low temperatures, the bulk mobility is limited by impurity scattering. It depends, as expected, on the doping level, and it also decreases with a decrease of temperature. In the modulation-doped channel, its mobility at temperatures above ≈ 80 K is comparable to the value of a lowly doped bulk sample. However, mobility is much enhanced at lower temperatures. The mod-

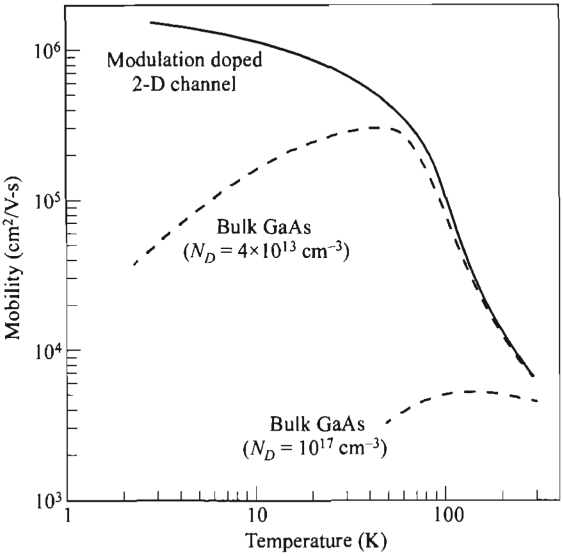


Fig. 17 Comparison of low-field electron mobility of modulation-doped 2-D channel to bulk GaAs at different doping levels. (After Ref. 36.)

ulation-doped channel does not suffer from impurity scattering which dominates at low temperatures. This benefit stems from the screening effect of a two-dimensional electron gas, where its conduction path is confined to a small cross-section which is smaller than 10 nm with high volume density.[33]

7.3.1 Basic Device Structure

The most-common heterojunctions for the MODFETs are the AlGaAs/GaAs, AlGaAs/InGaAs, and InAlAs/InGaAs heterointerfaces. A basic MODFET structure based on the AlGaAs/GaAs system is shown in Fig. 18. It is seen here that the barrier layer AlGaAs under the gate is doped, while the channel layer GaAs is undoped. This

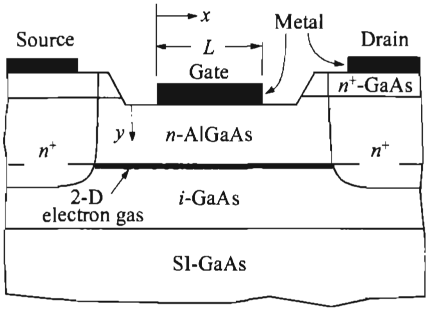


Fig. 18 Typical structure of MODFET, using the basic AlGaAs/GaAs system.

is the principle of modulation doping such that carriers from the doped barrier layer are transferred to reside at the heterointerface and are away from the doped region to avoid impurity scattering. The doped barrier layer is typically around 30-nm thick. Very often, a δ-doped charge sheet is used within the barrier layer and placed close to the channel interface, instead of uniform doping. The top layer of n^+-GaAs is for better source and drain ohmic contacts. These contacts are made from alloys containing Ge, such as AuGe. The source/drain deeper n^+-regions are formed either by ion implantation or introduced during the alloying step. Similar to MESFETs, the metal gate is sometimes made into the shape of a T-gate to reduce the gate resistance. Most MODFETs reported are n-channel devices for higher electron mobility.

7.3.2 *I-V* Characteristics

Based on the principle of modulation doping, the impurities within the barrier layer are completely ionized and carriers depleted away. Referring to the energy-band diagrams of Fig. 19, the potential variation within the depleted region is given by (see Section 2.2.3 on p. 88)

$$\psi_P = \frac{q}{\varepsilon_s}\int_0^{y_o} N_D(y)y\,dy \tag{72}$$

for a general doping profile. For uniform doping, this built-in potential becomes

$$\psi_P = \frac{qN_D y_o^2}{2\varepsilon_s}. \tag{73}$$

For a planar-doped charge sheet n_{sh} located at a distance of y_1 from the gate, this expression yields

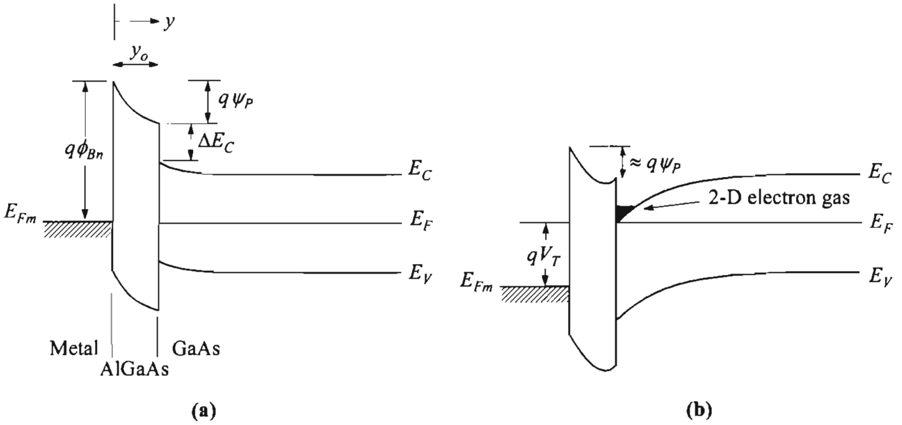


Fig. 19 Energy-band diagrams for an enhancement-mode MODFET at (a) equilibrium and (b) onset of threshold.

$$\psi_P = \frac{qn_{sh}y_1}{\varepsilon_s}. \tag{74}$$

The advantage here, compared to the uniformly doped AlGaAs layer, is the reduction of traps that are believed to be responsible for the anomalous behavior of current collapse at low temperature. The close proximity of dopants to the channel also gives a lower threshold voltage.

Like any other field-effect transistor, an important parameter is the threshold voltage, the gate bias at which the channel starts to form between the source and drain. From Fig. 19b, first-order approximation shows that this occurs when the Fermi level E_F at the GaAs surface coincides with the conduction-band edge E_C. This corresponds to the bias condition of:

$$V_T \approx \phi_{Bn} - \psi_P - \frac{\Delta E_C}{q}. \tag{75}$$

It can be seen here that by choosing the doping profile and barrier height ψ_{Bn}, V_T can be varied between positive and negative values. The example shown in Fig. 19 has a positive V_T, and the transistor is called an enhancement-mode (normally-off) device, as opposed to a depletion-mode (normally-on) device.

Once the threshold voltage is known, the rest of the analysis in deriving the *I-V* characteristics are similar to that for the MOSFETs. In getting to the final results here we skip some of the intermediate steps, and readers are referred to the MOSFET chapter for more-detailed derivations if necessary.

With gate voltage larger than the threshold voltage, the charge sheet in the channel induced by the gate is capacitively coupled and is given by

$$Q_n = C_o(V_G - V_T), \tag{76}$$

where

$$C_o = \frac{\varepsilon_s}{y_o + \Delta y} \tag{77}$$

and Δy is the channel thickness of the two-dimensional electron gas, estimated to be around 8 nm. When a drain bias is applied, the channel has a variable potential with distance and its value with respect to the source is designated as $\Delta\psi(x)$. It varies along the channel from 0 at the source to V_D at the drain. The channel charge as a function of position becomes

$$Q_n(x) = C_o[V_G - V_T - \Delta\psi(x)]. \tag{78}$$

The channel current at any location is given by

$$I_D(x) = ZQ_n(x)v(x). \tag{79}$$

Since the current is constant through out the channel, integrating the above equation from source to drain gives

$$I_D = \frac{Z}{L}\int_0^L Q_n(x)v(x)dx. \tag{80}$$

As for the other FETs, we derive the current equations with different assumptions on the velocity-field relationships.

Constant Mobility. With constant mobility, the drift velocity is simply given by

$$\begin{aligned} v(x) &= \mu_n \mathscr{E}(x) \\ &= \mu_n \frac{d\Delta\psi}{dx}. \end{aligned} \tag{81}$$

Substituting Eq. 81 into Eq. 80 and with proper change of variable, we obtain

$$I_D = \frac{Z\mu_n C_o}{L}\left[(V_G - V_T)V_D - \frac{V_D^2}{2}\right]. \tag{82}$$

The output characteristics for an enhancement-mode MODFET are shown in Fig. 20. In the linear region where $V_D \ll (V_G - V_T)$, Eq. 82 is reduced to an ohmic relationship,

$$I_{D\text{lin}} = \frac{Z\mu_n C_o(V_G - V_T)V_D}{L}. \tag{83}$$

At high V_D, $Q_n(L)$ at the drain is reduced to zero (Eq. 78), corresponding to the pinch-off condition, and current saturates with V_D. This gives a saturation drain bias of

$$V_{D\text{sat}} = V_G - V_T \tag{84}$$

and a saturation drain current of

$$I_{D\text{sat}} = \frac{Z\mu_n C_o}{2L}(V_G - V_T)^2. \tag{85}$$

From the above equation, the transconductance can be obtained,

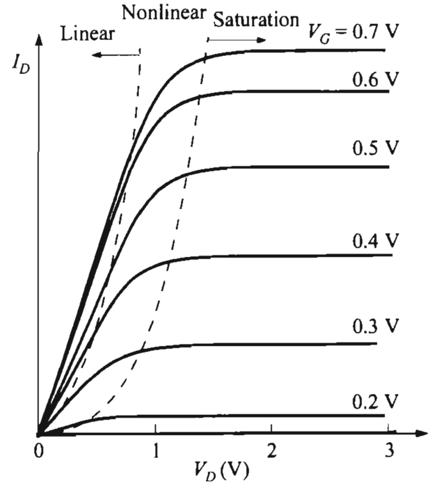


Fig. 20 Output characteristics of an enhancement-mode MODFET.

$$g_m \equiv \frac{dI_{Dsat}}{dV_G} = \frac{Z\mu_n C_o(V_G - V_T)}{L}. \tag{86}$$

Field-Dependent Mobility. In state-of-the-art devices, current becomes saturated with V_D before the pinch-off condition occurs, due to the fact that carrier drift velocity no longer is linearly proportional to the electric field. In other words, in high fields, the mobility becomes field dependent. For devices with high mobilities such as MODFETs, this phenomenon is more severe. Figure 21 shows the electron velocity-field relationship where a two-piece linear approximation is also shown with a critical field $\mathscr{E}_c$. Low-field electron mobilities reported for the AlGaAs/GaAs heterointerface are typically $\approx 10^4$ cm^2/V-s at 300 K, $\approx 2\times10^5$ cm^2/V-s at 77 K, and $\approx 2\times10^6$ cm^2/V-s at 4 K. The mobility enhancement at low temperatures in a MODFET is very pronounced as discussed before. But the improvement of υ_s at low temperatures is much less, ranging from 30 to 100%. High mobility also implies low $\mathscr{E}_c$, and the drain bias needed to drive the device towards velocity saturation is reduced. From the MOSFET equations, we set $M = 1$ (p. 306) since the channel doping is very light. Equations 36 and 37 in the MOSFET chapter (Chapter 6) become

$$I_{Dsat} = \frac{ZC_o\mu_n}{L}\left(V_G - V_T - \frac{V_{Dsat}}{2}\right)V_{Dsat}, \tag{87}$$

$$V_{Dsat} = L\mathscr{E}_c + (V_G - V_T) - \sqrt{(L\mathscr{E}_c)^2 + (V_G - V_T)^2}. \tag{88}$$

Velocity Saturation. In the case of short-channel devices, complete velocity saturation is approached and simpler equations can be used. The saturation current in this regime becomes

$$\begin{aligned} I_{Dsat} &= ZQ_n\upsilon_s \\ &= ZC_o(V_G - V_T)\upsilon_s, \end{aligned} \tag{89}$$

with the transconductance of

$$g_m \equiv \frac{dI_{Dsat}}{dV_G} = ZC_o\upsilon_s. \tag{90}$$

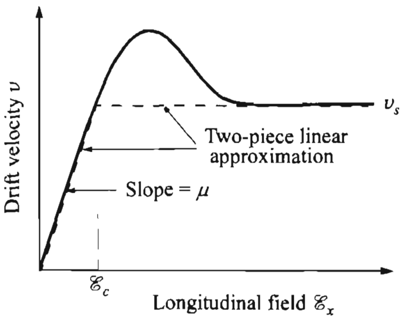


Fig. 21 υ-$\mathscr{E}$ relationship for the channel charge. Transfer-electron effect is shown for materials such as GaAs. Two-piece linear approximation is indicated.

Notice that in this extreme case $I_{D\text{sat}}$ is independent of L and g_m is independent of L and V_G.

At large V_G, the g_m, as indicated in Fig. 20, starts to decrease. The AlGaAs/GaAs heterointerface can confine a maximum carrier density Q_n/q of $\approx 1\times10^{12}$ cm^{-2}. Above this V_G ($1\times10^{12}q/C_o \approx 0.8$ V), charge is induced within the AlGaAs layer, whose mobility is much lower.

7.3.3 Equivalent Circuit and Microwave Performance

For discussions on small-signal equivalent circuit, f_T, $f_{\max}$, and noise, we can follow the analysis either in MOSFET or MESFET/JFET in the earlier part of this chapter. From the equivalent circuit, in the presence of parasitic source resistance, the extrinsic transconductance is degraded by

$$g_{mx} = \frac{g_m}{1+R_S g_m}. \tag{91}$$

The cutoff frequency f_T and the maximum frequency of oscillation $f_{\max}$ are given by,

$$f_T = \frac{g_m}{2\pi\left[C'_G\left(1+\dfrac{R_D+R_S}{R_{DS}}\right)+C'_{GD}g_m(R_D+R_S)+C'_{par}\right]} \approx \frac{g_m}{2\pi(ZLC_o+C'_{par})}, \tag{92}$$

$$f_{\max} \approx \sqrt{\frac{f_T}{8\pi R_G C'_{GD}}}. \tag{93}$$

The minimum noise figure is given as[33]

$$F_{\min} \approx 1 + 2\pi C_2 f C'_{GS}\sqrt{\frac{R_G+R_S}{g_m}} \tag{94}$$

where $C_2 = 1.6$ s/F, a lower noise figure compared to the corresponding value of 2.5 s/F for GaAs MESFETs (Eq. 71). Note that since C'_{GS} is proportional to L, devices with shorter channels have better noise performance.

The speed of MODFETs is higher than that of MESFETs, due to the higher mobilities. Even though the saturation velocities of these devices are comparable, higher mobility pushes the device toward the performance limit of complete velocity saturation. So for the same channel length, higher mobility always gives somewhat higher current and transconductance. Some examples of analog applications are low-noise small-signal amplifiers, power amplifiers, oscillators, and mixers. For digital applications, there are high-speed logic and RAM circuits. MODFET also has superior noise performance compared to other FETs. This improved noise property comes from higher current and transconductance.

Compared to the MESFET, the MODFET can support higher gate bias due to the additional barrier of the AlGaAs layer. It also has better scaling capability since it does not have the restriction associated with the channel depth ($L/a \geq \pi$, Eq. 70). Another advantage is lower-voltage operation because of the low $\mathscr{E}_c$ needed to drive the device into velocity saturation. One drawback of the MODFET is the limit of maximum charge-sheet density at the heterointerface which limits the maximum current drive.

We have pointed out before that the difference between a MODFET and a HIGFET is the presence of dopants in the barrier layer. It is interesting then to see the advantage of introducing these dopants in the barrier layer. Figure 22 compares the energy-band diagrams of these two devices. The comparison is made at the condition of equal channel charge or channel current, at whatever gate bias necessary. Note that at such condition, the region from the channel to the right side of the devices are identical. The difference lies in the barrier layer and to its left. One can observe that the doping in the barrier layer has two main functions. First, the threshold voltage is lowered. Second, the built-in potential within the barrier layer ψ_P increases the total barrier for carrier confinement. This higher barrier enables a higher gate bias before excessive gate current takes place.

7.3.4 Advanced Device Structures

The major development effort for MODFETs has been on a channel material that can further improve the electron mobility. Instead of GaAs, $In_xGa_{1-x}As$ has been pursued due to its smaller effective mass. Additional advantages include a larger ΔE_C because of the smaller bandgap. Its higher satellite band has less transfer-electron effect that degrades the mobility. These advantages are found to be directly related to the indium contents: the higher the percentage, the higher the performance.

Introduction of indium in InGaAs causes lattice mismatch to the GaAs substrate (see Fig. 32 on p. 56). However, growth of good-quality heteroepitaxial layer is still possible provided the epitaxial-layer thickness is under the so-called critical thickness (see discussion on p. 57), in which condition the epitaxial layer is under strain. Such technique yields a pseudomorphic InGaAs channel layer and the device is called pseudomorphic MODFET (P-MODFET or P-HEMT). Figure 23 summarizes the

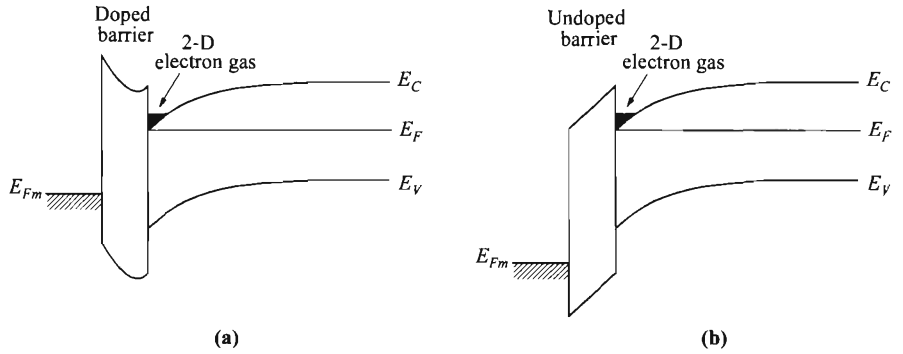


Fig. 22 Comparison of (a) doped barrier layer (as in MODFET) and (b) undoped barrier layer (as in HIGFET), for the same amount of channel charge.

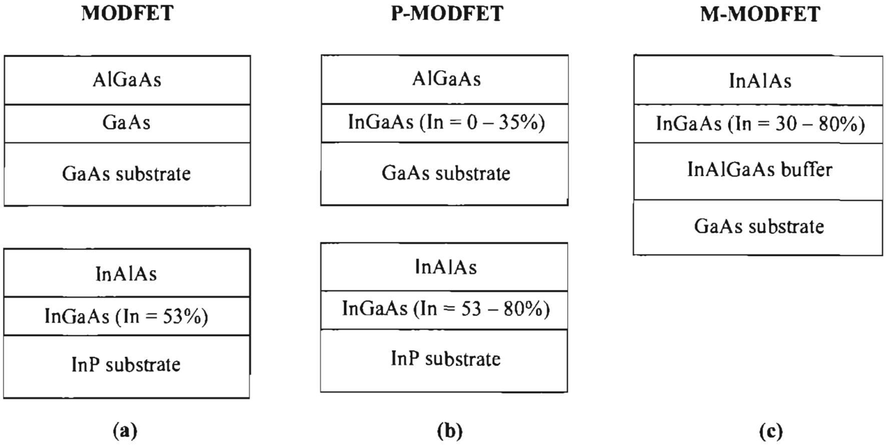


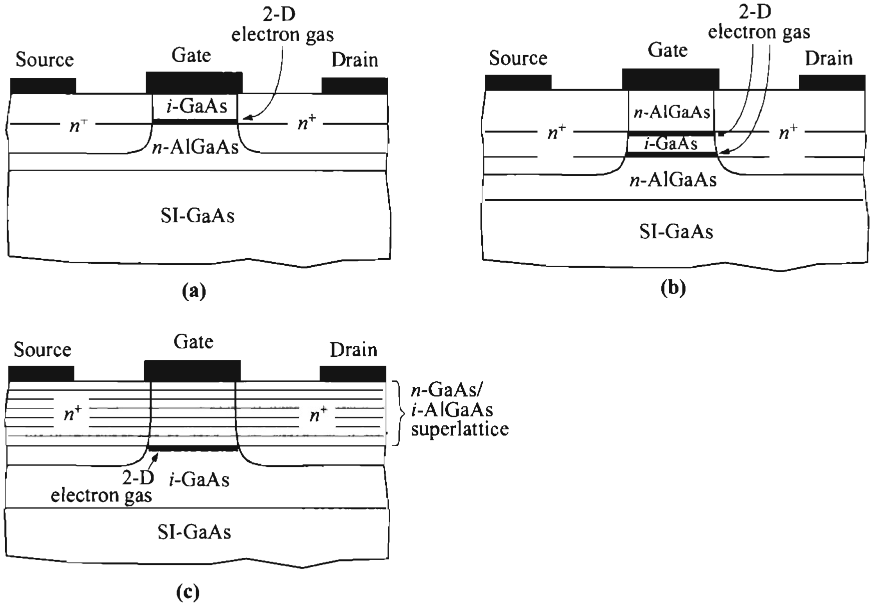
Fig. 23 Structures of (a) conventional unstrained MODFETs on GaAs and InP substrates, (b) pseudomorphic MODFETs (P-MODFETs), and (c) metamorphic MODFET (M-MODFET). Indium concentrations are indicated.

In% range for conventional MODFETs and P-MODFETs, on both GaAs and InP substrates. On GaAs substrate, P-MODFETs can accommodate a maximum of 35% indium. On InP substrate, an unstrained conventional MODFET starts with 53% In, and its P-MODFETs can contain as high as 80% In. So MODFET performance on InP substrate is always higher. The penalty is the higher cost of the InP substrate. In addition, an InP substrate is more susceptible to breakage during processing, and the wafer size is also smaller. These contribute to even higher cost. In general, P-MODFETs are sensitive to changes in strain during processing. Thermal budget has to be minimized to prevent relaxation of the pseudomorphic layer and introduction of dislocations that reduce the carrier mobility.

Yet another approach, depicted in Fig. 23c, is the latest innovation to obtain high In% on GaAs substrate. In this scheme, a thick buffer layer of graded composition is grown on the GaAs substrate. This thick buffer layer serves to transform the lattice constant gradually, from that of the GaAs substrate to whatever required for the subsequent growth of the InGaAs channel layer. In doing so, all the dislocations are contained within the buffer layer. The InGaAs channel layer is unstrained and dislocation-free. Such technique has permitted indium as high as 80%. The resultant MODFET is called metamorphic MODFET (M-MODFET).

Another material system for MODFET that has attracted increased interest recently is based on the AlGaN/GaN heterojunction. GaN has high energy gap (3.4 eV) and is attractive for power applications because of a low ionization coefficient and thus high breakdown voltage.[37] One interesting feature of the AlGaN/GaN system is the additional carriers coming from the effects of spontaneous polarization and piezoelectric polarization, apart from the modulation doping, resulting in higher current capability. In some cases, the AlGaN barrier layer is undoped and excess carrier concentration relies on these polarization effects.

To conclude this section, we show some variations in device structures that have certain advantages. Figure. 24a shows the inverted MODFET structure where the gate is placed over the channel layer rather than the barrier layer which is now grown over the substrate directly. In modulation doping, the high-E_g layer thickness determines the built-in potential ψ_P (Eq. 72) and preferably cannot be too thin. The channel layer does not have this restriction and can be thinner than the barrier layer. This gives a higher gate capacitance and thus higher transconductance and f_T. Another advantage is improved source and drain contact resistance since the contacts do not have to go through the high-E_g layer. The quantum-well MODFET, sometimes called a double-heterojunction MODFET, is shown in Fig. 24b. Because there are two parallel heterointerfaces, the maximum charge sheet and current are doubled. Another advantage is that the channel is sandwiched by two barriers, and the carriers have better confinement. Multiple-quantum-well structures also have been fabricated based on this principle. In the superlattice MODFET, the superlattice is used as the barrier layer (Fig. 24c). Within the superlattice, the narrow-E_g layers are doped while the wider-E_g layers are undoped. This structure eliminates traps in the AlGaAs layer, and also the parallel conduction path within this doped AlGaAs layer.


2-D
electron gas
Source
Gate
Drain
i-GaAs
n^+
n^+
n-AlGaAs
SI-GaAs
2-D
electron gas
Source
Gate
Drain
n-AlGaAs
n^+
n^+
i-GaAs
n-AlGaAs
SI-GaAs
Source
Gate
Drain
n^+
n^+
n-GaAs/
i-AlGaAs
superlattice
2-D
electron gas
i-GaAs
SI-GaAs

(a) (b) (c)

Fig. 24 Some variations of MODFET structures: (a) Inverted MODFET. (b) Quantum-well MODFET. (c) Superlattice MODFET.

REFERENCES

1. W. Shockley, "A Unipolar Field-Effect Transistor," *Proc. IRE,* **40**, 1365 (1952).
2. G. C. Dacey and I. M. Ross, "Unipolar Field-Effect Transistor," *Proc. IRE,* **41**, 970 (1953).
3. G. C. Dacey and I. M. Ross, "The Field-Effect Transistor," *Bell Syst. Tech. J.*, **34**, 1149 (1955).
4. C. A. Mead, "Schottky Barrier Gate Field-Effect Transistor," *Proc. IEEE*, **54**, 307 (1966).
5. W. W. Hooper and W. I. Lehrer, "An Epitaxial GaAs Field-Effect Transistor," *Proc. IEEE*, **55**, 1237 (1967).
6. K. Lehovec and R. Zuleez, "Voltage-Current Characteristics of GaAs JFETs in the Hot Electron Range," *Solid-State Electron.*, **13**, 1415 (1970).
7. R. E. Williams and D. W. Shaw, "Graded Channel FET's Improved Linearity and Noise Figure," *IEEE Trans. Electron Dev.*, **ED-25**, 600 (1978).
8. J. Ruch, "Electron Dynamics in Short Channel Field Effect Transistors," *IEEE Trans. Electron Dev.*, **ED-19**, 652 (1972).
9. K. Lehovec and R. Miller, "Field Distribution in Junction Field Effect Transistors at Large Drain Voltages," *IEEE Trans. Electron Dev.*, **ED-22**, 273 (1975).
10. H. Mizuta, K. Yamaguchi, and S. Takahashi, "Surface Potential Effect on Gate-Drain Avalanche Breakdown in GaAs MESFET's," *IEEE Trans. Electron Dev.*, **ED-34**, 2027 (1987).
11. R. J. Trew and U. K. Mishra, "Gate Breakdown in MESFET's and HEMT's," *IEEE Electron Dev. Lett.*, **EDL-12**, 524 (1991).
12. A. W. Ludikhuize, "A Review of RESURF Technology," *Proc. 12th Int. Symp. Power Semiconductor Devices & ICs*, p.11, 2000.
13. R. R. Bockemuehl, "Analysis of Field-Effect Transistors with Arbitrary Charge Distribution," *IEEE Trans. Electron Dev.*, **ED-10**, 31 (1963).
14. R. E. Williams and D. W. Shaw, "GaAs FETs with Graded Channel Doping Profiles," *Electron. Lett.*, **13**, 408 (1977).
15. R. A. Pucel, "Profile Design for Distortion Reduction in Microwave Field-Effect Transistors," *Electron. Lett.*, **14**, 204 (1978).
16. W, Liu, *Fundamentals of III-V Devices: HBTs, MESFETs, and HFETs/HEMTs*, Wiley, New York, 1999.
17. T. J. Maloney and J. Frey, "Frequency Limits of GaAs and InP Field-Effect Transistors at 300 K and 77 K with Typical Active Layer Doping," *IEEE Trans. Electron Dev.*, **ED-23**, 519 (1976).
18. K. Yokoyama, M. Tomizawa, and A. Yoshii, "Scaled Performance for Submicron GaAs MESFET's," *IEEE Electron Dev. Lett.*, **EDL-6**, 536 (1985).
19. M. F. Abusaid and J. R. Hauser, "Calculations of High-Speed Performance for Submicrometer Ion-Implanted GaAs MESFET Devices," *IEEE Trans. Electron Dev.*, **ED-33**, 913 (1986).
20. L. J. Sevin, *Field Effect Transistors*, McGraw-Hill, New York, 1965.
21. R. J. Trew, "SiC and GaN Transistors—Is There One Winner for Microwave Power Applications?" *Proc. IEEE*, **90**, 1032 (2002).

22. J. Shealy, J. Smart, M. Poulton, R. Sadler, D. Grider, S. Gibb, B. Hosse, B. Sousa, D. Halchin, V. Steel, et al., "Gallium Nitride (GaN) HEMT's: Progress and Potential for Commercial Applications," *IEEE GaAs Integrated Circuits Symp.*, p. 243, 2002.
23. R. A. Pucel, H. A. Haus, and H. Statz, "Signal and Noise Properties of GaAs Microwave Field-Effect Transistors," in L. Martin, Ed., *Advances in Electronics and Electron Physics*, Vol. **38**, Academic, New York, p. 195, 1975.
24. H. Fukui, "Optimal Noise Figure of Microwave GaAs MESFETs," *IEEE Trans. Electron Dev.*, **ED-26**, 1032 (1979).
25. S. C. Binari, P. B. Klein, and T. E. Kazior, "Trapping Effects in GaN and SiC Microwave FETs," *Proc. IEEE*, **90**, 1048 (2002).
26. L. Esaki and R. Tsu, "Superlattice and Negative Conductivity in Semiconductors," *IBM Research*, RC 2418, March 1969.
27. R. Dingle, H. L. Stormer, A. C. Gossard, and W. Wiegmann, "Electron Mobilities in Modulation-Doped Semiconductor Heterojunction Superlattices," *Appl. Phys. Lett.*, **33**, 665 (1978).
28. H. L. Stormer, R. Dingle, A. C. Gossard, W. Wiegmann, and M. D. Sturge, "Two-Dimensional Electron Gas at a Semiconductor-Semiconductor Interface," *Solid State Commun.*, **29**, 705 (1979).
29. T. Mimura, S. Hiyamizu, T. Fujii, and K. Nanbu, "A New Field-Effect Transistor with Selectively Doped GaAs/n-Al$_x$Ga$_{1-x}$As Heterojunctions," *Jpn. J. Appl. Phys.*, **19**, L225 (1980).
30. T. Mimura, "The Early History of the High Electron Mobility Transistor (HEMT)," *IEEE Trans. Microwave Theory Tech.*, **50**, 780 (2002).
31. D. Delagebeaudeuf, P. Delescluse, P. Etienne, M. Laviron, J. Chaplart, and N. T. Linh, "Two-Dimensional Electron Gas M.E.S.F.E.T. Structure," *Electron. Lett.*, **16**, 667 (1980).
32. H. Daembkes, Ed., *Modulation-Doped Field-Effect Transistors: Principle, Design and Technology*, IEEE Press, Piscataway, New Jersey, 1991.
33. H. Morkoc, H. Unlu, and G. Ji, *Principles and Technology of MODFETs: Principles, Design and Technology*, vols. 1 and 2, Wiley, New York, 1991.
34. C. Y. Chang and F. Kai, *GaAs High-Speed Devices*, Wiley, New York, 1994.
35. M. Golio and D. M. Kingsriter, Eds, *RF and Microwave Semiconductor Devices Handbook*, CRC Press, Boca Raton, Florida, 2002.
36. P. H. Ladbrooke, "GaAs MESFETs and High Mobility Transistors (HEMT)," in H. Thomas, D. V. Morgan, B. Thomas, J. E. Aubrey, and G. B. Morgan, Eds., *Gallium Arsenide for Devices and Integrated Circuits*, Peregrinus, London,1986.
37. U. K. Mishra, P. Parikh, and Y. F. Wu, "AlGaN/GaN HEMTs—An Overview of Device Operation and Applications," *Proc. IEEE*, **90**, 1022 (2002).

PROBLEMS

1. For a JFET with a power-law doping $N = N_{D2}x^n$ where N_{D2} and n are constants. Find I_D vs. V_G and g_m when $n \rightarrow \infty$.

2. An n-channel GaAs MESFET has been fabricated on semiinsulating substrate. It has a uniformly doped channel of $N_D = 10^{17}$ cm^{-3}, with $\phi_{Bn} = 0.9$ V, $a = 0.2$ µm, $L = 1$ µm, and $Z = 10$ µm.
   (a) Is this an enhancement- or depletion-mode device?
   (b) Find the threshold voltage.
   (c) Find the saturation current at $V_G = 0$ (for constant mobility of 5,000 cm^2/V-s).
3. Derive Eq. 19 by substituting ψ_{bi} in Eq. 15 to Eq. 17.
4. Design a GaAs MESFET with a maximum transconductance of 200 mS/mm and a drain saturation current $I_{D\text{sat}}$ of 200 mA/mm at zero gate-source bias. Assume $I_{D\text{sat}} = \beta(V_G - V_T)^2$ and $\beta \equiv Z\mu\varepsilon_s/2aL$, $\mu = 5{,}000$ cm^2/V-s, $L = 1$ µm, and $\psi_{bi} = 0.6$ V.
5. Show that (a) for a MESFET the measured drain conductance in the linear region is given by $g_{D0}/[1+(R_S+R_D)g_{D0}]$, and (b) the measured transconductance in the saturation region is given by $g_m/(1+R_S g_m)$ where R_S and R_D are the source and drain resistance, respectively.
6. An InP MESFET has $N_D = 2\times10^{17}$ cm^{-3}; $L = 1.5$ µm, $L/a = 5$, and $Z = 75$ µm. Assume $v_s = 6\times10^6$ cm/s, $\psi_{bi} = 0.7$ V. Using the saturated-velocity model, find the cutoff frequency for $V_G = -1$ V and $V_D = 0.2$ V (at which the channel near the drain is just pinched off).
7. For very-large-scale integrated circuits, the maximum allowed power per MESFET gate is 0.5 mW. Assume a clock frequency of 5 GHz and a node capacitor of 32 fF, find the upper limit of V_{DD} (in volts).
8. An InP MESFET has $N_D = 10^{17}$ cm^{-3}, $L = 1.5$ µm, $a = 0.3$ µm, $Z = 75$ µm. Assume $v_s = 6\times10^6$ cm/s, $\psi_{bi} = 0.7$ V, applied gate voltage = −1 V, and $\varepsilon_s = 12.4\varepsilon_o$. From the saturation-velocity model, find the cutoff frequency.
9. Find the thickness of the undoped spacer layer d_s, such that the two-dimensional electron gas concentration of an AlGaAs/GaAs heterojunction is 1.25×10^{12} cm^{-2} at zero gate bias. Assume that in the n-AlGaAs, the first 50 nm is doped to 1×10^{18} cm^{-3}, and the remaining layer of thickness d_s is undoped. The Schottky barrier height is 0.89 V, $\Delta E_C/q = 0.23$ V, and the dielectric constant of AlGaAs is 12.3.
10. (a) Find the threshold voltages of a conventional and a delta-doped heterostructure AlGaAs-GaAs FETs.
    (b) Evaluate the variations of these threshold voltages for two-monolayer fluctuations in AlGaAs layer thickness.
    Assuming that one monolayer ≈ 3 Å in AlGaAs, the Schottky barrier height is 0.9 V, the conduction-band discontinuity is 0.3 eV, the uniform doping in the conventional HEFT is 10^{18} cm^{-3} with a thickness of 40 nm, the delta doping is located 40 nm from the metal-semiconductor interface, with a sheet charge density of 1.5×10^{12} cm^{-2}, and the dielectric permittivity for AlGaAs is assumed to be 10^{-12} F/cm.
11. In an AlGaAs/GaAs MODFET, the n-type $Al_{0.3}Ga_{0.7}As$ layer is doped to 10^{18} cm^{-3}, and has a thickness of 50 nm. Assume an undoped spacer layer of 2 nm, a barrier height of 0.85 eV, and a conduction-band discontinuity of 0.22 eV. The dielectric constant for the ternary is 12.2. Find the two-dimensional electron concentration at the source for $V_G = 0$.
12. Consider an AlGaAs/GaAs MODFET with a 50-nm AlGaAs and 10-nm undoped AlGaAs spacer. Assume the threshold voltage is −1.3 V, $N_D = 5\times10^{17}$ cm^{-3}, $\Delta E_C = 0.25$ eV, the channel width is 8 nm, and the dielectric constant is 12.3. Calculate the Schottky barrier height and the 2-D electron gas concentration at $V_G = 0$.

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

PART IV

NEGATIVE-RESISTANCE AND POWER DEVICES

- **Chapter 8** **Tunnel Devices**
- **Chapter 9** **IMPATT Diodes**
- **Chapter 10** **Transferred-Electron and Real-Space-Transfer Devices**
- **Chapter 11** **Thyristors and Power Devices**

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

8

Tunnel Devices

8.1 INTRODUCTION
8.2 TUNNEL DIODE
8.3 RELATED TUNNEL DEVICES
8.4 RESONANT-TUNNELING DIODE

8.1 INTRODUCTION

In this chapter we consider devices based on quantum-mechanical tunneling. In the classical sense, carriers having energy smaller than some potential barrier height are confined or stopped by the barrier completely. In quantum mechanics, the wave nature of carriers are considered, and the wave does not terminate abruptly at the boundary of the barrier. As a result, not only can the carriers have finite probability of existence inside the barrier, they can leak through the barrier if the barrier width is thin enough. This leads to the concept of tunneling probability and tunneling current. Basic tunneling phenomenon has been discussed in Section 1.5.7 where tunneling probability is introduced.

The tunneling process and devices based on this phenomenon have some interesting properties. First, the tunneling phenomenon is a majority-carrier effect, and the tunneling time of carriers through the potential energy barrier is not governed by the conventional transit time concept ($\tau = W/v$, where W is the barrier width and v is the carrier velocity), but rather by the quantum transition probability per unit time which is proportional to $\exp[-2\langle k(0)\rangle W]$, where $\langle k(0)\rangle$ is the average value of momentum encountered in the tunneling path corresponding to an incident carrier with zero transverse momentum and energy equal to the Fermi energy.[1] Reciprocation gives the tunneling time proportional to $\exp[2\langle k(0)\rangle W]$. This tunneling time is very short, permitting the use of tunnel devices well into the millimeter-wave region. Secondly, since the tunneling probability depends on the available states of both the originating side and the receiving side, tunneling current is not monotonically dependent on the bias, and negative differential resistance can result.

A perceived drawback of tunnel devices might be the low current density allowed, but in fact tunnel devices can have substantial current densities exceeding 1.5 mA/μm^2 in SiGe interband tunnel diodes[2] and 4.5 mA/μm^2 in InP-based resonant-tunneling diodes.[3] For this reason, investigation of integrated tunnel diode and transistor circuits has continued, especially to enable power reduction through the use of more efficient circuit topologies.[4]

In this chapter, the two main tunnel devices considered are the *tunnel diode* and the *resonant-tunneling diode*. Originally when it was first discovered, the tunnel diode seemed to have great potential. Time has shown that its application in the real market has been very limited. This is due to the difficulty in fabrication and reproducibility, especially if incorporated in integrated circuits because abrupt and high doping profiles are called for. The tunnel diode is now being replaced by the Gunn diode and the IMPATT diode as oscillators, and by FETs as switching elements. The more recent resonant-tunneling diode brings up another form of tunneling that is fundamentally interesting. The phenomenon of resonant tunneling has also been incorporated in many other devices and one example will be given at the end of the chapter.

8.2 TUNNEL DIODE

The tunnel diode was discovered by L. Esaki in 1958 and is often called the Esaki diode.[5] As part of his Ph.D. dissertation work, Esaki was studying heavily doped germanium *p-n* junctions for application in high-speed bipolar transistors in which a narrow and heavily doped base was required.[6–7] He discovered an *anomalous* current-voltage characteristic in the forward direction, that is, a negative-differential-resistance region (negative dI/dV) over part of the forward characteristics. Esaki explained this anomalous characteristic by the quantum tunneling concept and obtained reasonable agreement between the tunneling theory and the experimental results. Subsequently tunnel diodes were demonstrated by researchers on other semiconductor materials, such as GaAs[8] and InSb[9] in 1960, Si[10] and InAs[11] in 1961, and GaSb[12] and InP[13] in 1962.

A tunnel diode consists of a simple *p-n* junction in which both *p*- and *n*-sides are degenerate (i.e., very heavily doped with impurities) and in sharp transition. Figure 1 shows a schematic energy diagram of a tunnel diode in thermal equilibrium. Because of the high dopings the Fermi levels are located within the allowed bands. The amount of degeneracy, V_p' and V_n', is typically a few kT/q, and the depletion-layer width is of the order of 10 nm or less, which is considerably narrower than the conventional *p-n* junction. [In this chapter we use V_n' and V_p' ($= -V_n$ and $-V_p$) to give positive values and to be consistent with notations in other chapters.]

Figure 2a shows typical static current-voltage characteristics of a tunnel diode. In the reverse direction (*p*-side negative bias with respect to *n*-side) the current increases monotonically. In the forward direction the current first increases to a maximum

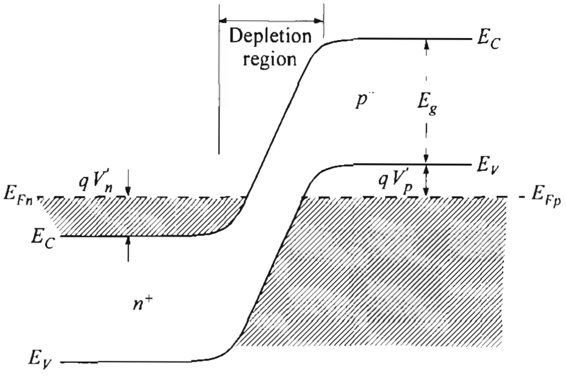


Fig. 1 Energy-band diagram of tunnel diode in thermal equilibrium. V_p' and V_n' are degeneracies on the *p*-side and *n*-side respectively.

value (peak current or I_P) at a peak voltage V_P, then decreases to a minimum value I_V at a valley voltage V_V. For voltages much larger than V_V, the current increases exponentially with the voltage. The static characteristics are the result of three current components: band-to-band tunneling current, excess current, and diffusion current (Fig. 2b).

We first discuss qualitatively the tunneling processes at absolute zero temperature, using the simplified band structure in Fig. 3 which shows band alignment of the *p*- and *n*-sides when a bias is applied.[14] The corresponding current is also designated

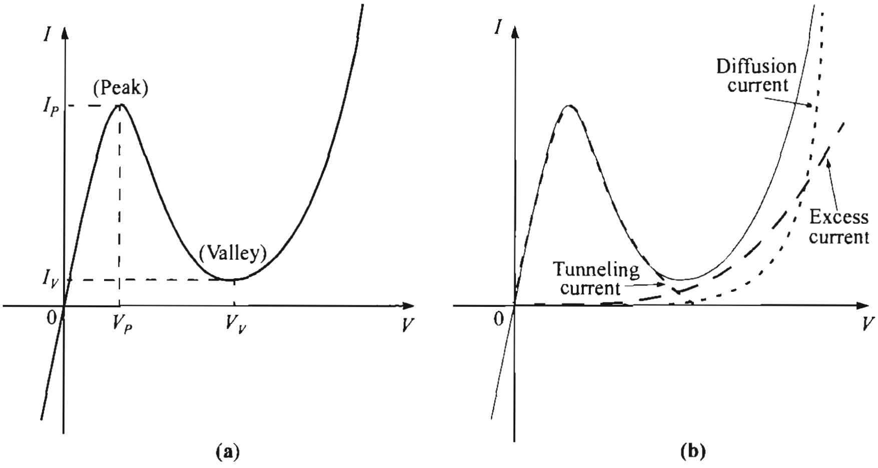


Fig. 2 (a) Static current-voltage characteristics of a typical tunnel diode. I_P and V_P are the peak current and peak voltage. I_V and V_V are the valley current and valley voltage. (b) The total static characteristics are broken down into three current components.

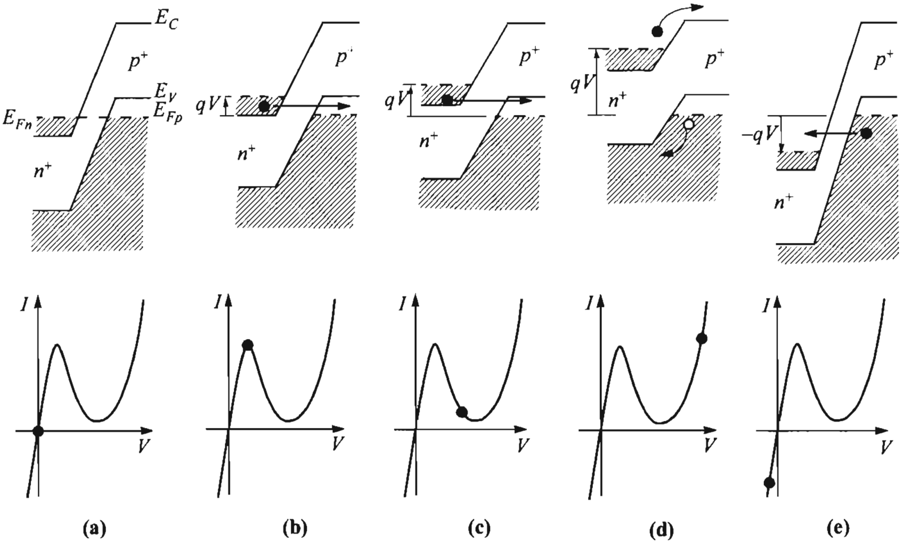


Fig. 3 Simplified energy-band diagrams of tunnel diode at (a) thermal equilibrium, zero bias; (b) forward bias V such that peak current is obtained; (c) forward bias approaching valley current; (d) forward bias with diffusion current and no tunneling current; and (e) reverse bias with increasing tunneling current. (After Ref. 14.)

by the dot on the I-V curve. Note that the Fermi levels are within the allowed bands of the semiconductor, and at thermal equilibrium (Fig. 3a) the Fermi level is constant across the junction. Above the Fermi level there are no filled states (electrons) on either side of the junction, and below the Fermi level there are no empty states (holes) available on either side of the junction. Hence, the net tunneling currents at zero applied voltage is zero.

When a voltage is applied, the electrons may tunnel from the conduction band to the valence band, or vice versa. The necessary conditions for tunneling are: (1) occupied energy states exist on the side from which the electron tunnels; (2) unoccupied energy states exist at the same energy level on the side to which the electron can tunnel; (3) the tunneling potential barrier height is low and the barrier width is small enough that there is a finite tunneling probability; and (4) the momentum is conserved in the tunneling process.

When a forward bias is applied (Fig. 3b), there exists a common band of energies in which there are filled states on the n-side and unoccupied states on the p-side. The electrons can thus tunnel from the n-side to the p-side and energy is conserved. When the forward voltage is further increased, this band of common energies decreases (Fig. 3c). If forward voltage is applied such that the bands are *uncrossed*, that is, the edge of the n-type conduction band is exactly opposite to the top of the p-type valence band, there are no available states opposite to filled states. Thus at this point and onward the tunneling current can no longer flow. With still further increase of the voltage the normal diffusion current and excess current start to dominate (Fig. 3d).

One thus expects that as the forward voltage increases from zero, the tunneling current increases from zero to a maximum I_P and then decreases to zero when $V = V'_n + V'_p$, where V is the applied forward voltage, V'_n the amount of degeneracy on the n-side [$V'_n \equiv (E_{Fn} - E_C)/q$], and V'_p is the amount of degeneracy on the p-side [$V'_p \equiv (E_V - E_{Fp})/q$], as shown in Fig. 1. The decreasing current after the peak gives rise to the negative differential resistance. The Fermi levels for degenerate semiconductors are inside the conduction band or valance band, and they are given by (see Section 1.4.1);[15]

$$qV'_n \equiv E_F - E_C \approx kT\left[\ln\left(\frac{n}{N_C}\right) + 2^{-3/2}\left(\frac{n}{N_C}\right)\right], \tag{1a}$$

$$qV'_p \equiv E_V - E_F \approx kT\left[\ln\left(\frac{p}{N_V}\right) + 2^{-3/2}\left(\frac{p}{N_V}\right)\right], \tag{1b}$$

where m_{de} and m_{dh} are density-of-state effective masses for electrons and holes.

Figure 3e shows electron tunneling from the valence band into the conduction band when a reverse bias is applied. In this direction, the tunneling current increases with bias indefinitely, and there is no negative differential resistance.

The tunneling process can be either direct or indirect, and these are demonstrated in Fig. 4 where the E-k relationships are superimposed on the classical turning points of the tunnel junction. Figure 4a shows direct tunneling when the electrons can tunnel from the vicinity of the conduction-band minimum to the vicinity of the valence-band maximum, and at the same time without a change of momentum (in k-space). For direct tunneling to occur, the conduction-band minimum and the valence-band maximum must have the same momentum. This condition can be fulfilled by semiconductors that have a direct bandgap, such as GaAs and GaSb. This condition can also be fulfilled by semiconductors with indirect bandgap, such as Si and Ge, when the applied voltage is sufficiently large that electrons tunnel from the higher direct conduction-band minimum (Γ point) rather than from the lower satellite minimum.[16]

Indirect tunneling occurs in semiconductors of indirect bandgap, i.e., the conduction-band minimum does not align at the same momentum as the valence-band maximum (Fig. 4b) in the E-k relationship. To conserve momentum, the difference in momentum between the conduction-band minimum and the valence-band maximum must be supplied by scattering agents such as phonons or impurities. For phonon-assisted tunneling, both their energy and momentum must be conserved; that is, the sum of the phonon energy and the initial electron energy is equal to the final electron energy after it has tunneled, and the sum of the initial electron momentum and the phonon momentum ($\hbar k_p$) is equal to the final electron momentum after it has tunneled. In general, the probability for indirect tunneling is much lower than the probability for direct tunneling when direct tunneling is possible. Also, indirect tunneling involving several phonons has a much lower probability than that with only a single phonon.

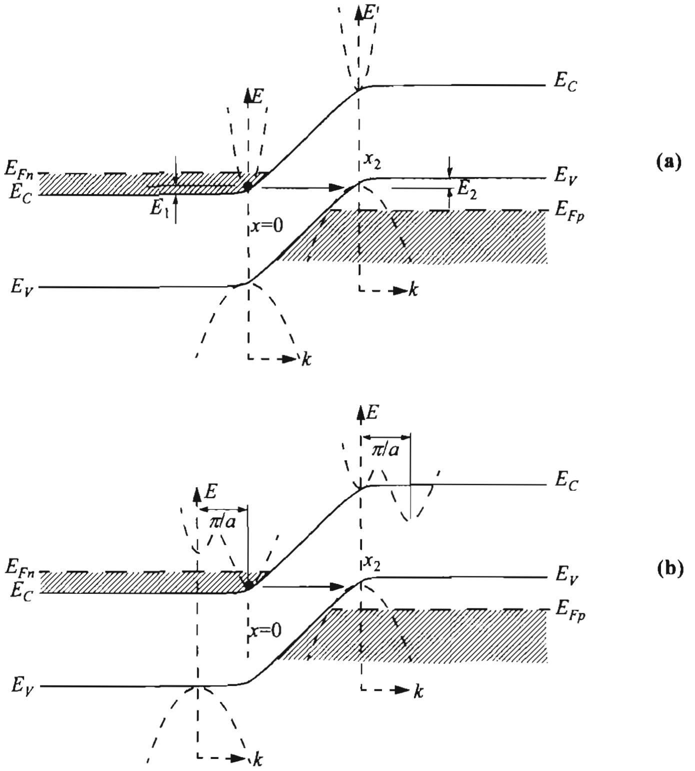


Fig. 4 Direct and indirect tunneling processes demonstrated by E-k relationship superimposed on the classical turning points ($x = 0$ and x_2) of the tunnel junction. (a) Direct tunneling process with $k_{\text{min}} = k_{\text{max}}$. (b) Indirect tunneling process with $k_{\text{min}} \neq k_{\text{max}}$.

8.2.1 Tunneling Probability and Tunneling Current

In this section we focus on the component of the tunneling current. When the electric field in a semiconductor is sufficiently high, on the order of 10^6 V/cm, a finite probability exists for interband quantum tunneling, i.e., direct transition of electrons from the conduction band into the valence band, or vice versa. The tunneling probability T_t (see Section 1.5.7) can be given by the WKB (Wentzel-Kramers-Brillouin) approximation[17]

$$T_t \approx \exp\left[-2\int_0^{x_2} |k(x)|\, dx\right] \tag{2}$$

where $|k(x)|$ is the absolute value of the wave vector of the carrier inside the barrier, and $x = 0$ and x_2 are the classical boundaries shown in Fig. 4.

The tunneling of an electron through a forbidden band is formally the same as a particle tunneling through a potential barrier. Examination of Fig. 4 indicates that the tunnel barrier is of the triangular shape that is shown explicitly in Fig. 5. We start with the general equation for the E-k relationship

$$k(x) = \sqrt{\frac{2m^*}{\hbar^2}(PE - E_C)} \tag{3}$$

where PE is the potential energy. For tunneling consideration, the incoming electron has an PE equal to the bottom of the energy gap. The value inside the square root is thus negative and k is imaginary. Furthermore, the varying conduction-band edge E_C can be expressed in terms of the electric field $\mathscr{E}$. The wave vector inside the triangular barrier is given by

$$k(x) = \sqrt{\frac{2m^*}{\hbar^2}(-q\mathscr{E}x)}. \tag{4}$$

Substituting Eq. 4 into Eq. 2 yields

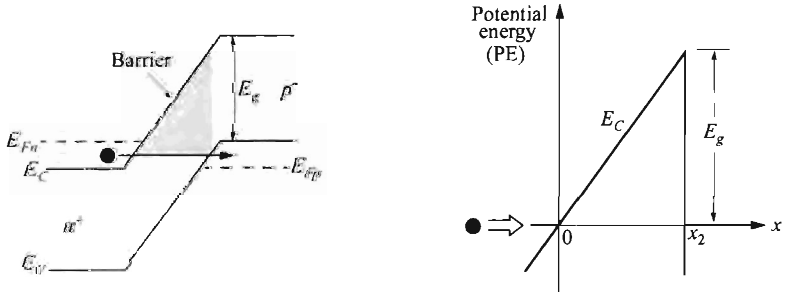
$$T_t \approx \exp\left[-2\int_0^{x_2}\sqrt{\frac{2m^*}{\hbar^2}(q\mathscr{E}x)}dx\right]. \tag{5}$$

Since for a triangular barrier with a uniform field, $x_2 = E_g/\mathscr{E}q$, we have the result

$$T_t \approx \exp\left(-\frac{4\sqrt{2m^*}E_g^{3/2}}{3q\hbar\mathscr{E}}\right). \tag{6}$$

From the result it is clear that to obtain large tunneling probability, both the effective mass and the bandgap should be small and the electric field should be large.

We next proceed to calculate the tunneling current and shall present the first-order approach using the density of states in the conduction band and valence band. We shall also assume direct tunneling where the momentum is conserved in direct bandgap. At thermal equilibrium, the tunneling current $I_{C\to V}$ from the conduction band to the empty states of the valence band and the current $I_{V\to C}$ from the valence



(a) (b)

Fig. 5 (a) Tunneling in a tunnel diode can be analyzed by (b) a triangular potential barrier.

band to the empty states of the conduction band should be balanced. Expressions for $I_{C\to V}$ and $I_{V\to C}$ are formulated as follows:

$$I_{C\to V} = C_1\int F_C(E)N_c(E)T_t[1-F_V(E)]N_v(E)dE, \tag{7a}$$

$$I_{V\to C} = C_1\int F_V(E)N_v(E)T_t[1-F_C(E)]N_c(E)dE, \tag{7b}$$

where C_1 is a constant, the tunneling probability T_t is assumed to be equal for both directions, $F_C(E)$ and $F_V(E)$ are the Fermi-Dirac distribution functions, and $N_c(E)$ and $N_v(E)$ are the density of states in the conduction band and valence band, respectively. When the junction is forward biased, the net tunneling current I_t is given by

$$I_t = I_{C\to V}-I_{V\to C} = C_1\int_{E_{Cn}}^{E_{Vp}}[F_C(E)-F_V(E)]T_tN_c(E)N_v(E)dE. \tag{8}$$

Note that the limits of integration are from E_C of the n-side (E_{Cn}) to E_V of the p-side (E_{Vp}). Rigorous manipulation of Eq. 8 leads to the following result[18]

$$J_t = \frac{q^2\mathscr{E}}{36\pi\hbar^2}\sqrt{\frac{2m^*}{E_g}}D\exp\left(-\frac{4\sqrt{2m^*}E_g^{3/2}}{3q\hbar\mathscr{E}}\right), \tag{9}$$

where the integral D is

$$D\equiv\int[F_C(E)-F_V(E)]\left[1-\exp\left(-\frac{2E_S}{\bar{E}}\right)\right]dE, \tag{10}$$

and the average electric field is given by

$$\mathscr{E} = \sqrt{\frac{q(\psi_{bi}-V)N_AN_D}{2\varepsilon_s(N_A+N_D)}}, \tag{11}$$

where ψ_{bi} is the built-in potential. In Eq. 10, E_S is the smaller of E_1 and E_2 (Fig. 4a), and $\bar{E}$ is given by

$$\bar{E}\equiv\frac{\sqrt{2}q\hbar\mathscr{E}}{\pi\sqrt{m^*E_g}}. \tag{12}$$

For a Ge tunnel diode, the appropriate effective mass in Eq. 9 is given by[19]

$$m^* = 2\left(\frac{1}{m_e^*}+\frac{1}{m_{lh}^*}\right)^{-1} \tag{13}$$

for tunneling from the light-hole band to the ⟨000⟩ conduction band, where m_{lh}^* is the light-hole mass (= $0.044m_o$) and m_e^* is the ⟨000⟩ conduction-band mass (= $0.036m_o$). For tunneling in the ⟨100⟩ direction to the ⟨111⟩ minima, the effective mass is given by

$$m^* = 2\left[\left(\frac{1}{3m_l^*}+\frac{2}{3m_t^*}\right)+\frac{1}{m_{lh}^*}\right]^{-1} \tag{14}$$

where $m_l^* = 1.6m_o$ and $m_t^* = 0.082m_o$ are the longitudinal and transverse masses of the $\langle 111 \rangle$ minima. The exponents in Eq. 9 differ, however, by only 5% in these two cases.

The quantity D, Eq. 10, is an overlap integral which modulates the shape of the I-V curve. It has the dimensions of energy and it depends on the temperature and the degeneracy V_n' and V_p'. At $T = 0$ K, both F_C and F_V are step functions. Figure 6 shows the quantity D in relative scale versus the forward voltage for the case $V_n' > V_p'$. The drop of D to zero corresponds to the valley voltage and it occurs at

$$V_V = V_n' + V_p' . \tag{15}$$

The prefactor in Eq. 9 gives an idea on the magnitude of the tunneling current. Figure 7 plots the peak current calculated from Eq. 9 for several Ge tunnel diodes, together with experimental values that show very good agreement.

It is difficult to get the whole I-V characteristics of the tunneling current because the analytical solution of Eq. 9 is complicated. However, the tunneling current has been found to fit an empirical formula quite well, given in the form of;

$$I_t = \frac{I_P V}{V_P} \exp\left(1 - \frac{V}{V_P}\right), \tag{16}$$

where I_P and V_P are the peak current and peak voltage as defined in Fig. 2. Knowing the peak current, the critical parameter remaining is the peak voltage. This peak voltage can be obtained by a different approach. In this approach, we find the carrier profiles of the electrons in the conduction band in the n-side, and holes in the valence band in the p-side. Under bias, when the peaks of these two profiles line up at the same energy, this is the peak voltage for the peak tunneling current. This concept is demonstrated in Fig. 8.

The carrier profile is given by the product of occupancy and density of states. For electrons and holes, they are given by;

$$n(E) = F_C(E) N_c(E), \tag{17a}$$

$$p(E) = [1 - F_V(E)] N_v . \tag{17b}$$

For degenerate n-type semiconductor, the electron profile can be written as[20]

$$n(E) = \frac{8\pi (m^*)^{3/2} \sqrt{2(E - E_C)}}{h^3 \{1 + \exp[(E - E_F)/kT]\}} . \tag{18}$$

The energy for the peak concentration can be obtained by differentiating Eq. 18 with respect to E. The resultant equations cannot be solved explicitly, but it has been shown that with good approximation, the energy for maximum electron density occurs at the energy level of[20]

$$E_{mn} = E_{Fn} - \frac{qV_n'}{3} . \tag{19a}$$

A similar approach and result can be obtained for the p-side;

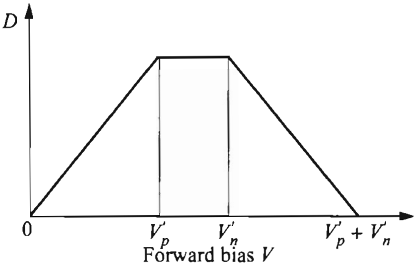
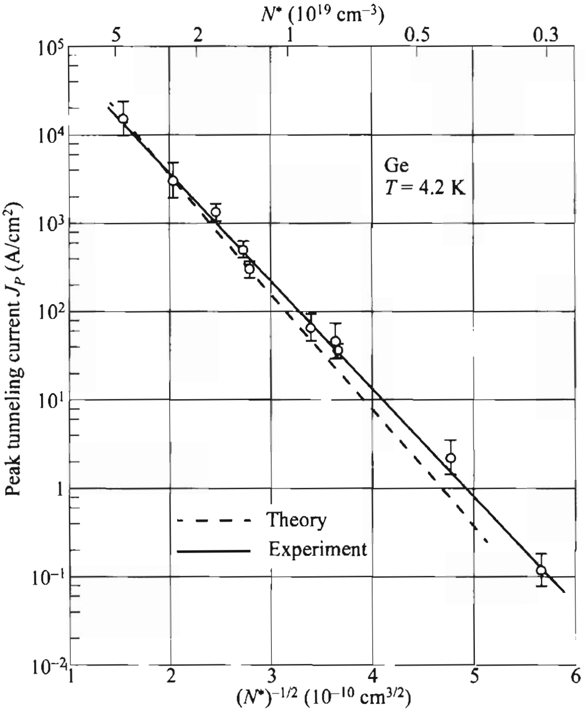


Fig. 6 Integral D (relative scale) vs. forward voltage V, for direct tunneling with small $\bar{E}$ and $V'_n > V'_p$. (After Ref. 18.)



$$N^* \equiv \frac{N_A N_D}{N_A + N_D}$$

Fig. 7 Peak tunneling current density vs. effective doping concentration of Ge tunnel diodes. The dashed line is calculated from Eq. 9. (After Refs. 20 and 21.)

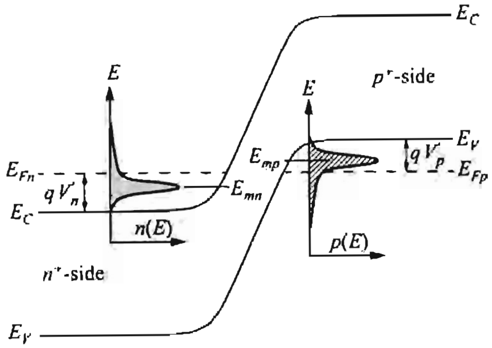

E_C
E
p^+-side
E
E_V
qV'_p
E_{Fp}
E_{mp}
E_{Fn}
qV'_n
E_{mn}
E_C
$n(E)$
$p(E)$
n^+-side
E_V

Fig. 8 Density profiles for electrons and holes in n-type and p-type degenerate semiconductors. E_{mn} and E_{mp} are their peak energies.

$$E_{mp} = E_{Fp} + \frac{qV'_p}{3}. \tag{19b}$$

The peak voltage is simply the bias necessary to align these two peak energies, given by

$$\begin{aligned} V_P &= (E_{mp} - E_{mn})/q \\ &= \frac{V'_n + V'_p}{3}. \end{aligned} \tag{20}$$

Figure 9 shows the position of the peak voltage as a function of the degeneracy V'_n and V'_p for Ge tunnel diodes. Note that the peak voltage shifts toward higher values as the doping increases. The experimental values of V_P agree reasonably well with Eq. 20.

Up to now we have not considered the requirement of the conservation of momentum. This could have two effects both of which will reduce the tunneling probability and tunneling current. The first effect is indirect tunneling for indirect bandgap materials where the change of momentum in k-space has to be compensated by some scattering effects, such as phonon scattering and impurity scattering. For phonon-assisted indirect tunneling, the tunneling probability is reduced by a multiplier to Eq. 6, except that in Eq. 6, E_g is to be replaced by $E_g + E_p$ where E_p is the phonon energy.[18,22] The expression for the tunneling current is similar in form to Eq. 9 but its magnitude is much lower. So the readers are reminded that for indirect tunneling, the equations presented in this chapter have to be modified.

The second effect associated with momentum is its vector direction in relationship to the direction of tunneling. In previous discussion all the kinetic energy is assumed to be in the direction of tunneling. In reality, we have to divide the total energy into

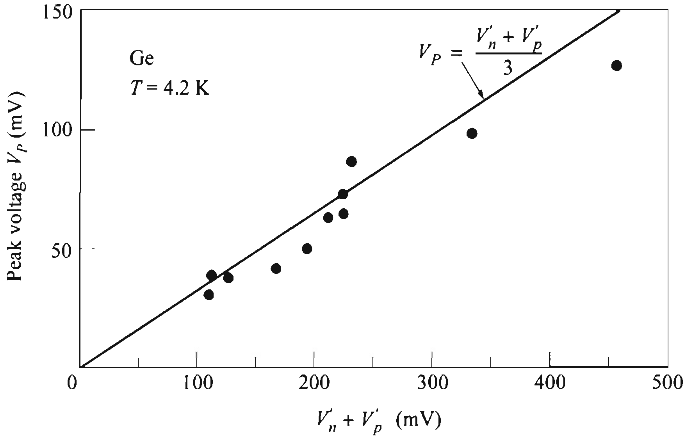


Fig. 9 Variation of peak voltage of Ge tunnel diodes as a function of the sum $V'_n + V'_p$. (After Refs. 20 and 21.)

E_x and $E_\perp$, where $E_\perp$ is the energy associated with momentum perpendicular to the direction of tunneling (or the transverse momentum) and E_x is the energy associated with momentum in the tunneling direction;

$$E = E_x + E_\perp = \frac{\hbar^2 k_x^2}{2m_x^*} + \frac{\hbar^2 k_\perp^2}{2m_\perp^*}, \tag{21}$$

where the subscripts x and $\perp$ designate their components in the parallel and perpendicular directions to the tunneling. Considering that only the component E_x contributes to the tunneling process, the tunneling probability is reduced by the amount of $E_\perp$, to the value of

$$T_t \approx \exp\left(-\frac{4\sqrt{2m^*}E_g^{3/2}}{3q\hbar\mathscr{E}}\right)\exp\left(-\frac{E_\perp\pi\sqrt{2m^*E_g}}{q\hbar\mathscr{E}}\right). \tag{22}$$

In other words, perpendicular energy further reduces the transmission by the factor of the second exponential term, a measure of the transverse momentum.

8.2.2 Current-Voltage Characteristics

As shown in Fig. 2b, the static I-V characteristic is the result of three current components: the tunneling current, the excess current, and the diffusion current. For an ideal tunnel diode, the tunneling current decreases to zero at biases where $V \geq (V_n' + V_p')$; for larger biases only normal diode currents caused by forward injection of minority carriers flow. In practice, however, the actual current at such biases is considerably in excess of the normal diode current, hence the term *excess current*. The excess current is mainly due to carrier tunneling by way of energy states within the forbidden gap.

The excess current is derived with the help of Fig. 10, where some examples of possible tunneling routes are shown.[10] An electron could drop down from C to an empty level at B, from which it could tunnel to D (route CBD). Alternatively, the electron starting at C in the conduction band might tunnel to an appropriate local level at

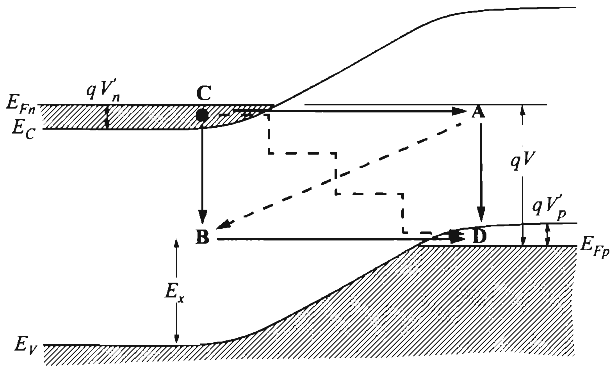


Fig. 10 Band diagram illustrating mechanisms of tunneling via states in the forbidden gap for the excess current. (After Ref. 10.)

A, from which it could then drop down to the valence band *D* (route *CAD*). A third variant is a route such as *CABD*, where the electron dissipates its excess energy in a process that could be called impurity-band conduction between *A* and *B*. A fourth route that should also be included is a staircase from *C* to *D*. It consists of a series of tunneling transitions between local levels together with a series of vertical steps in which the electron loses energy by transferring from one level to another, a process made possible when the concentration of intermediate levels is sufficiently high. The first route *CBD* can be regarded as the basic mechanism, while the other routes are simply more-complicated modifications.

Let the junction be at a forward bias V, and consider an electron making a tunneling transition from *B* to *D*. The energy E_x through which it must tunnel is given by

$$\begin{aligned} E_x &\approx E_g + q(V_n' + V_p') - qV \\ &\approx q(\psi_{bi} - V) \end{aligned} \tag{23}$$

where ψ_{bi} is the built-in potential. The tunneling probability T_t for the electron on the level at *B* can be given by an expression the same as Eq. 6

$$T_t \approx \exp\left(-\frac{4\sqrt{2m_x^*}E_x^{3/2}}{3q\hbar\mathscr{E}}\right) \tag{24}$$

except here E_g is replaced by E_x and the appropriate mass m_x^* should be used. Furthermore, let the volume density of the occupied levels at *B* be D_x. Then the excess current density will be given by

$$J_x \approx C_2 D_x T_t \tag{25}$$

where C_2 is a constant. It is assumed that the excess current will vary predominantly with the parameters in the exponent of T_t rather than with those in the factor D_x. Substituting Eqs. 23, 24, 11 into Eq. 25 yields the expression for the excess current:[10]

$$J_x \approx C_2 D_x \exp\{-C_3[E_g + q(V_n' + V_p') - qV]\} \tag{26}$$

where C_3 is another constant. Equation 26 predicts that the excess current will increase with the volume density of bandgap levels (through D_x), and also increase exponentially with the applied voltage V (provided that $qV \ll E_g$). Equation 26 can also be rewritten as[23]

$$J_x = J_V \exp[C_4(V - V_V)] \tag{27}$$

where J_V is the valley current density at the valley voltage V_V and C_4 is the prefactor in the exponent. Experimental results of In(J_x) versus V for common tunnel diodes exhibit linear relationships in good agreement with Eq. 27. Note that there is no negative differential resistance in this type of tunneling.

The diffusion current is the familiar minority-carrier injection current in *p*-*n* junctions:

$$J_d = J_0\left[\exp\left(\frac{qV}{kT}\right) - 1\right] \tag{28}$$

where J_0 is the saturation current density given in Chapter 2 (Eq. 64). The complete static current-voltage characteristic is the sum of the three current components:

$$\begin{aligned} J &= J_t + J_x + J_d \\ &= \frac{J_P V}{V_P}\exp\left(1 - \frac{V}{V_P}\right) + J_V \exp[C_4(V - V_V)] + J_0 \exp\left(\frac{qV}{kT}\right). \end{aligned} \tag{29}$$

Each component dominates the current in some voltage range. The tunneling current's contribution to the total current is significant for $V < V_V$, the excess current's contribution is significant for $V \approx V_V$, and the contribution of the diffusion current is significant for $V > V_V$.

Figure 11 shows a comparison of the typical current-voltage characteristics of Ge, GaSb, and GaAs tunnel diodes at room temperature. The current ratios of I_P/I_V are 8:1 for Ge, 12:1 for GaSb,[24] and 28:1 for GaAs.[25] Tunnel diodes have been made in many other semiconductors, such as Si with a current ratio of about 4:1.[26] The ultimate limitation on the ratio depends on (1) the peak current, which depends on the dopings, effective tunneling mass, and the bandgap; and (2) the valley current, which depends on the distribution and concentration of energy levels in the forbidden gap. So the ratio for a given semiconductor can be increased by increasing the doping concentrations on both n- and p-sides, increasing the sharpness of their profiles, and minimizing defect densities.

We shall briefly consider the I-V characteristics resulting from the effects of temperature, electron bombardment, and pressure. The temperature variation of the peak current can be explained by the change of the integral D and E_g in Eq. 9. At high concentrations the temperature effect on D is small, and the negative value of dE_g/dT is primarily responsible for the change in tunneling probability. As a result, the peak current increases with temperature. In the more-lightly doped tunnel diodes, the decrease of D with temperature dominates, and the temperature coefficient is nega-

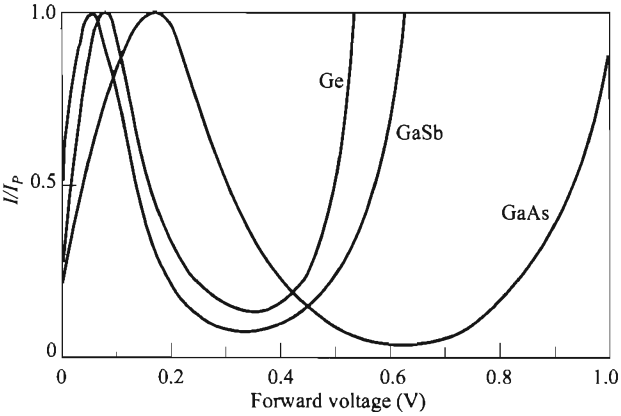


Fig. 11 Typical I-V characteristics of Ge, GaSb, and GaAs tunnel diodes at 300 K.

tive. For typical Ge tunnel diodes, the variation of the peak current over a temperature range of –50 to 100°C is about ±10%.[27] The valley current generally increases with increasing temperature, because the bandgap reduces with temperature.

After electron bombardment, the major effect is the increase in excess current caused by increased volume density of the energy levels in the bandgap.[28] The increased excess current can be annealed out gradually. Similar results can be observed for other types of radiation, such as γ rays. Physical stress on the device causes the excess current in Ge and Si tunnel diodes to increase.[29] The changes are found to be reversible. This effect arises from deep-lying states associated with the strain-induced defects in the depletion region. For GaSb, however, both I_P and I_V decrease with increasing hydrostatic pressure,[30] which can be explained by an increase in the bandgap and a reduction in the degeneracy of V_n' and V_p' with increasing pressure.

8.2.3 Device Performance

Originally, most tunnel diodes are made using one of the following techniques. (1) Ball alloy: A small metal alloy pellet containing the counter dopant of high-solid solubility is alloyed to the surface of a semiconductor substrate with high doping, in a precisely controlled temperature-time cycle under inert or hydrogen gas (e.g., the arsenic in an arsenic-doped tin ball forms the n^+-region on the surface of a p^+-Ge substrate). (2) Pulse bond: The contact and the junction are made simultaneously when the junction is pulse-formed between the semiconductor substrate and the metal alloy containing the counter dopant. (3) Planar processes:[31] Planar tunnel diode fabrication uses planar technology, including solution growth, diffusion, and controlled alloy. More-recent techniques are based on low-temperature epitaxial growth where the dopants are incorporated during the growth of the semiconductor layer. These include MBE (molecular-beam epitaxy) and MOCVD (metal-organic chemical vapor deposition). These techniques yield higher peak-to-valley ratios because of higher and sharper doping profiles for higher peak tunneling current, as well as lower defect densities for lower excess current.

Figure 12 shows the basic equivalent circuit, which consists of four elements: the series inductance L_S, the series resistance R_S, the diode capacitance C_j, and the negative diode resistance $-R$. The series resistance R_S includes the on-chip interconnects and external wire resistance, the ohmic contacts, and the spreading resistance in the wafer substrate, which is given by $\rho/2d$ where ρ is the resistivity of the semiconductor and d is the diameter of the diode area. The series inductance L_S is due to interconnects, wire bond, and external wires. We shall see that these parasitic elements establish important limits on the performance of the tunnel diode.

To consider the intrinsic diode capacitance and negative resistance, we refer to typical dc current-voltage characteristics in Fig. 13a. Figure 13b shows the conductance plot (dI/dV) versus bias. At the peak and valley voltages the conductance becomes zero. The diode capacitance is usually measured at the valley voltage, and is designated by C_j. The differential resistance, defined as $(dI/dV)^{-1}$, is plotted in Fig. 13c. The absolute value of the negative resistance at the inflection point, which

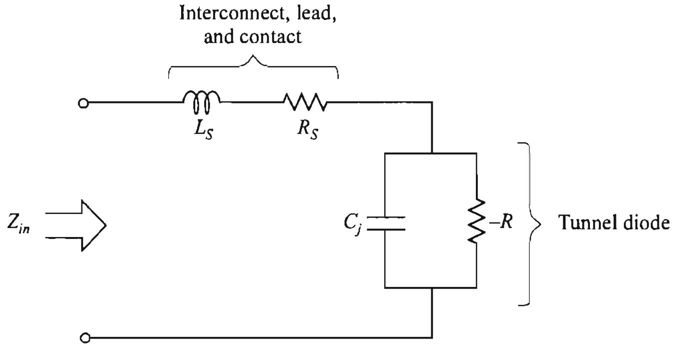

Interconnect, lead, and contact
L_S
R_S
Z_{in}
C_j
$-R$
Tunnel diode

Fig. 12 Equivalent circuit of tunnel diode.

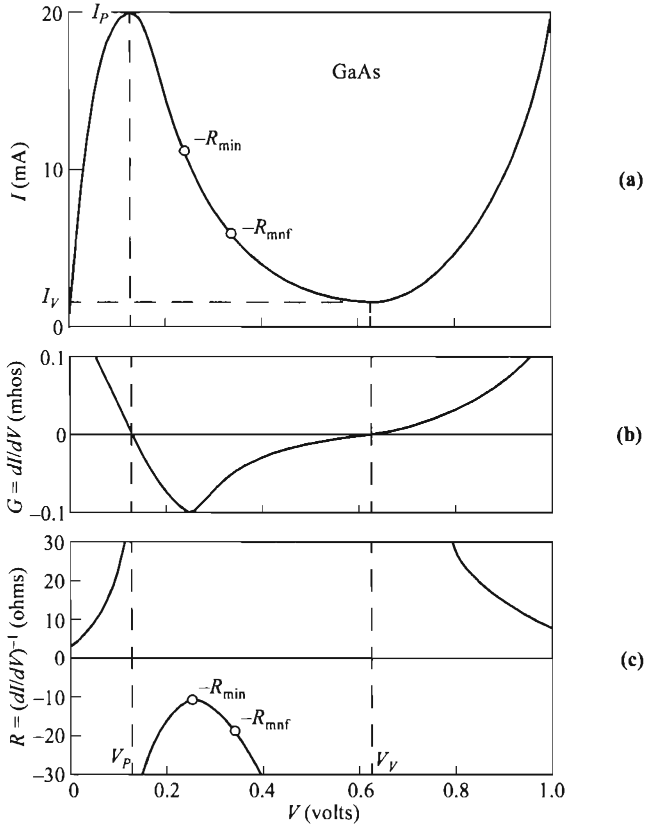


Fig. 13 (a) Intrinsic current-voltage characteristics of a GaAs tunnel diode at 300 K. (b) Differential conductance $G = dI/dV$ vs. V. At peak and valley currents, $G = 0$. (c) Differential resistance $(dI/dV)^{-1}$ vs. V, where R_{min} is the minimum negative resistance and R_{mnf} is the resistance corresponding to the minimum noise figure.

is the minimum negative resistance in the region, is designated by R_{min}, This resistance can be approximated by

$$R_{\text{min}} \approx \frac{2V_P}{I_P} \tag{30}$$

where V_P and I_P are the peak voltage and peak current, respectively.

The total input impedance Z_{in} of the equivalent circuit of Fig. 12 is given by

$$Z_{in} = \left[R_S + \frac{-R}{1+(\omega RC_j)^2}\right] + j\left[\omega L_S + \frac{-\omega C_j R^2}{1+(\omega RC_j)^2}\right]. \tag{31}$$

From Eq. 31 we see that the resistive (real) part of the impedance will be zero at a certain frequency, and the reactive (imaginary) part of the impedance will also be zero at another frequency. We denote these frequencies by the resistive cutoff frequency f_r and the reactive cutoff frequency f_x, respectively. These frequencies are given by

$$f_r = \frac{1}{2\pi RC_j}\sqrt{\frac{R}{R_S} - 1}, \tag{32}$$

$$f_x = \frac{1}{2\pi}\sqrt{\frac{1}{L_S C_j} - \frac{1}{(RC_j)^2}}. \tag{33}$$

Since R is bias dependant, so are the cutoff frequencies. These resistive and reactive cutoff frequencies specified at the bias of R_{min} are

$$f_{r0} \equiv \frac{1}{2R_{\text{min}}C_j}\sqrt{\frac{R_{\text{min}}}{R_S} - 1} \geq f_r \tag{34}$$

$$f_{x0} \equiv \frac{1}{2\pi}\sqrt{\frac{1}{L_S C_j} - \frac{1}{(R_{\text{min}}C_j)^2}} \leq f_x. \tag{35}$$

Since at that bias, the value of R is at its minimum (R_{min}), f_{r0} is the maximum resistive cutoff frequency at which the diode no longer exhibits net negative resistance; and f_{x0} is the minimum reactive cutoff frequency (or the self-resonant frequency) at which the diode reactance is zero. It follows that the diode would oscillate if $f_{r0} > f_{x0}$. In most applications where the diode is operated into the negative-resistance region, it is desirable to have $f_{x0} > f_{r0}$ and $f_{r0} \gg f_0$, f_0 being the operating frequency. Equations 34 and 35 show that to fulfill the requirement that $f_{x0} > f_{r0}$, the series inductance L_S must be low.

The switching speed of a tunnel diode is determined by the current available for charging the junction capacitance and the average RC product. Since R, the negative resistance, is inversely proportional to the peak current, a large tunneling current is required for fast switching. A figure of merit for tunnel diodes is the speed index, which is defined as the ratio of the peak current to the capacitance at the valley voltage, I_P/C_j. Figure 14 shows the speed index and the peak current versus depletion-layer width of Ge tunnel diodes at 300 K. We see that a narrow depletion width or large effective doping is needed to obtain a large speed index.

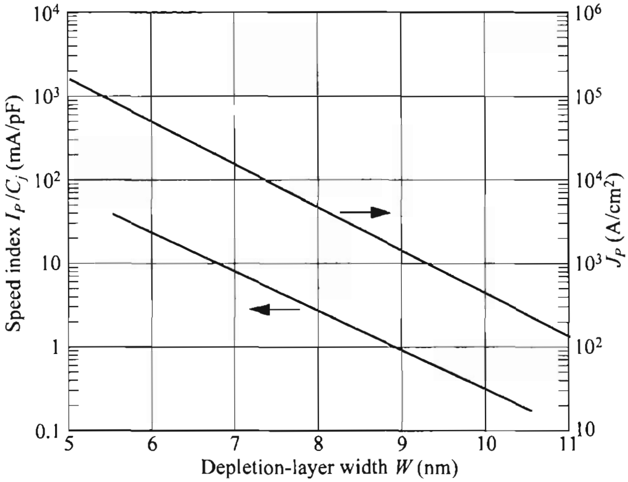


Fig. 14 Average value of the speed index ($= I_P/C_j$) and peak current density vs. depletion-layer width of Ge tunnel diodes at 300 K. (After Ref. 31.)

Another important quantity associated with the equivalent circuit is the noise figure, which is given as

$$NF = 1 + \frac{q}{2kT}|RI|_{\min} \tag{36}$$

where $|RI|_{\min}$ is the minimum value of the negative resistance-current product on the current-voltage characteristic. Figure 13 indicates the corresponding value of R (designated by R_{mnf}). The product $q|RI|_{\min}/2kT$ is called the noise constant and is a material constant. Typical values of the noise constant at room temperature are 1.2 for Ge, 2.4 for GaAs, and 0.9 for GaSb. The noise figure for Ge tunnel diodes is about 5–6 dB at around 10 GHz.

In addition to its microwave and digital applications, the tunnel diode is a useful device for the study of fundamental physical parameters. The diode can be used in tunneling spectroscopy, a technique that uses tunneling electrons of known energy distribution as a spectroscopic probe instead of photons of known frequency in optical spectroscopy. Tunneling spectroscopy has been used to study electron energy states in solids and to observe the excitation of modes. For example, from the shape of the I-V characteristics of a Si tunnel diode at low temperature, the phonon-assisted tunneling processes can be identified.[32] Similar observations are made in group III-V semiconductor junctions with the plots of the conductance (dI/dV) versus bias at 4.2 K for GaP, InAs, and InSb.[33,34]

8.3 RELATED TUNNEL DEVICES

8.3.1 Backward Diode

In connection with the tunnel diode, when the doping concentration on the p-side or n-side is nearly or not quite degenerate, the current in the *reverse* direction for small bias, as shown in Fig. 15, is larger than the current in the *forward* direction—hence the name *backward diode*. At thermal equilibrium, the Fermi level in the backward diode is very close to the band edges. When a small reverse bias (p-side negative with respect to n-side) is applied, the energy-band diagram is similar to Fig. 3e except that there is no degeneracy on both sides. Under reverse bias, electrons can readily tunnel from the valence band into the conduction band and give rise to a tunneling current, given by Eq. 9, which can be written in the form

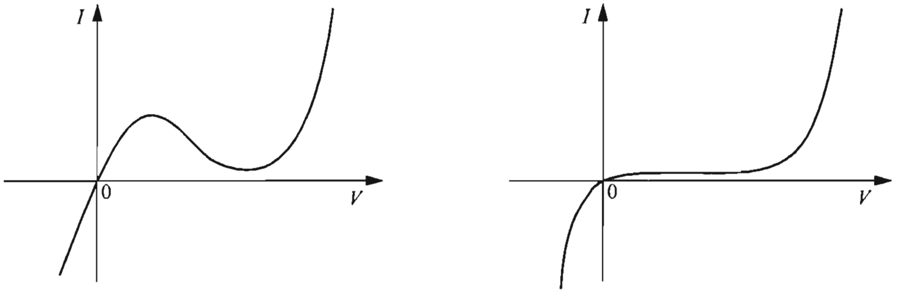
$$J \approx C_5 \exp\left(\frac{|V|}{C_6}\right) \tag{37}$$

where C_5 and C_6 are positive quantities and are slowly varying functions of the applied voltage V. Equation 37 indicates that the reverse current increases approximately exponentially with the voltage.

The backward diode can be used for rectification of small signals, and microwave detection and mixing.[35] Similar to the tunnel diode, the backward diode has a good frequency response because there is no minority-carrier storage.[36] In addition, the current-voltage characteristic is insensitive to temperature and to radiation effects, and the backward diode has very low $1/f$ noise.[37]

For nonlinear applications such as high-speed switching, a device figure of merit is γ, the ratio of the second derivative to the first derivative of the current-voltage characteristics. It is also referred to as the curvature coefficient:[38]

$$\gamma \equiv \frac{d^2I/dV^2}{dI/dV}. \tag{38}$$



(a) (b)

Fig. 15 Comparison of (a) tunnel diode with negative resistance and (b) backward diode without negative resistance.

The value of γ is a measure of the degree of nonlinearity normalized to the operating admittance level. For a regular forward-biased *p-n* junction or Schottky barrier the value of γ is simply given by q/nkT. Thus γ varies inversely with T. At room temperature, γ for an ideal *p-n* junction ($n = 1$) is about 40 V^{-1} independent of bias. For a reverse-biased *p-n* junction, however, the value of γ is very small at low voltages and increases linearly with the avalanche multiplication factor near breakdown voltage.[39] Although the reverse breakdown characteristic could in theory give a value of γ greater than 40 V^{-1}, because of the statistical distribution of impurities and the effect of space-charge resistance, much lower values of γ are expected.

For a backward diode the value of γ can be obtained from Eq. 16 and is given by[40]

$$\gamma(V=0) = \frac{4}{V_n' + V_p'} + \frac{2}{\hbar}\sqrt{\frac{2\varepsilon_s m^*(N_A + N_D)}{N_A N_D}} \tag{39}$$

where m^* is the average effective mass of the carriers

$$m^* \approx \frac{m_e^* m_h^*}{m_e^* + m_h^*}. \tag{40}$$

Clearly, the curvature coefficient γ depends on the impurity concentrations on both sides of the junction and the effective masses. In contrast to Schottky barriers, the value of γ is relatively insensitive to temperature variation because the parameters in Eq. 39 are slowly varying functions of temperature.

Figure 16 shows a comparison between theoretical and experimental values of γ for Ge backward diodes. The solid lines are computed from Eq. 39 using $m_e^* = 0.22m_0$ and $m_h^* = 0.39m_0$. The agreement is generally good over the doping range considered, and γ can exceed 40 V^{-1}.

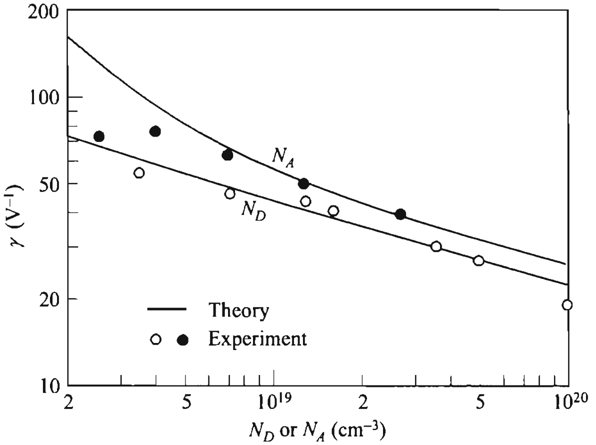


Fig. 16 Curvature coefficient vs. N_A (for a fixed $N_D = 2\times10^{19}$ cm^{-3}) or N_D (for a fixed $N_A = 10^{19}$ cm^{-3}), in Ge at 300 K and $V = 0$. (After Ref. 40.)

8.3.2 MIS Tunnel Devices

For a metal-insulator-semiconductor (MIS) structure, the current-voltage characteristics critically depend on the insulator thickness. If the insulator layer is sufficiently thick (greater than ≈ 7 nm for the Si-SiO$_2$ system), carrier transport through the insulator layer is negligible and the MIS structure represents a conventional MIS capacitor (discussed in Chapter 4). On the other hand, if the insulator layer is very thin (less than 1 nm), little impediment is met by carriers transporting between the metal and the semiconductor, and the behavior resembles a Schottky-barrier diode. Inbetween these two oxide thicknesses, there exist also different tunneling mechanisms. We will examine in more details, in particular, Fowler-Nordheim tunneling (Fig. 17a), direct tunneling (Fig. 17b), MIS tunnel diode with ultra-thin oxide (Fig. 17c), and finally negative resistance resulting from MIS tunnel diode on degenerate substrate.

Fowler-Nordheim Tunneling. Fowler-Nordheim (F-N) tunneling is characterized by (1) the barrier has a triangular shape, and (2) tunneling through only part of the insulator layer. It is shown in Fig. 17a that with a higher field, a narrower barrier is in effect. After tunneling through this triangular barrier, the rest of the insulator does not impede the current flow. So the total insulator layer only affects the current indirectly by affecting the field. The F-N current has the form similar to Eq. 9 and is given as[41]

$$J = \frac{q^2\mathscr{E}^2}{16\pi^2\hbar\phi_{ox}}\exp\left[\frac{-4\sqrt{2m^*}(q\phi_{ox})^{3/2}}{3\hbar q\mathscr{E}}\right] = C_4\mathscr{E}^2\exp\left(\frac{-C_5}{\mathscr{E}}\right) \tag{41}$$

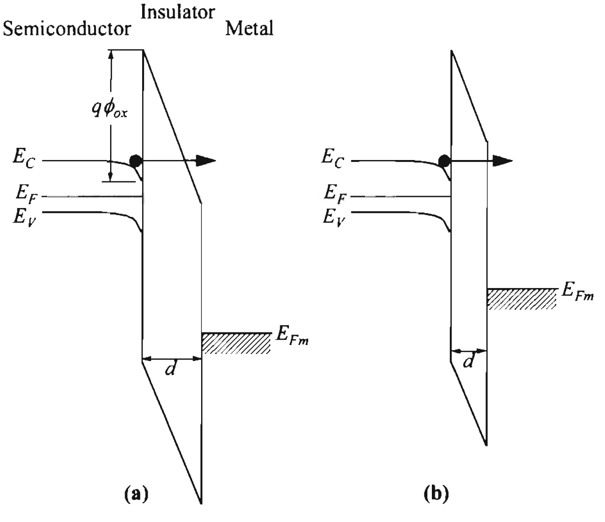


Fig. 17 Tunneling mechanism depends on oxide thickness range. (a) In thicker oxides (> 5 nm) Fowler-Nordheim tunneling through a triangular barrier and only part of the insulator layer. (b) Direct tunneling through whole insulator layer. (c) MIS tunnel diode (d < 30 Å) is characterized by nonequilibrium ($E_{Fn} \neq E_{Fp}$) and tunneling by both types of carriers.

where for thermal oxides, the constants are $C_4 = 9.63\times10^{-7}$ A/V^2 and $C_5 = 2.77\times10^8$ V/cm. The commonality between this equation and Eq. 9 is the triangular barrier. But in F-N tunneling, in the WKB approximation, the band structure in the insulator layer, including the effective mass, has to be used instead. Note that the insulator thickness is not in the formula but only the field. The transition between F-N tunneling and direct tunneling is demonstrated in Fig. 18. Direct tunneling occurs at thinner oxides and lower fields. The oxide thickness for transition between F-N tunneling and direct tunneling can be approximated by $d = \phi_{ox}/\mathscr{E}$. For electron tunneling, $\phi_{ox} = 3.1$ V, and $\mathscr{E}$ for medium tunneling current is around 6 MV/cm. This gives an oxide thickness of ≈ 5 nm.

Direct Tunneling. Direct tunneling occurs below ≈ 5 nm and with such a thin insulator, other phenomena such as quantum effects cannot be ignored. In quantum mechanics, the peak carrier concentration of the inversion layer is at a finite distance from the semiconductor-insulator interface, so the effective insulator thickness is increased. Furthermore, the inversion layer is a quantum well and carriers are at quantized energy levels above the conduction-band edge. With these quantum effects, a simple expression for direct current is not accurate. Simulated results are shown in Fig. 19. It can be seen that the tunneling current is very sensitive to the oxide thickness. Another factor to consider is that in practical devices such as MOSFETs, the top electrode on oxide is a heavily doped poly-Si layer instead of metal. Such a contact

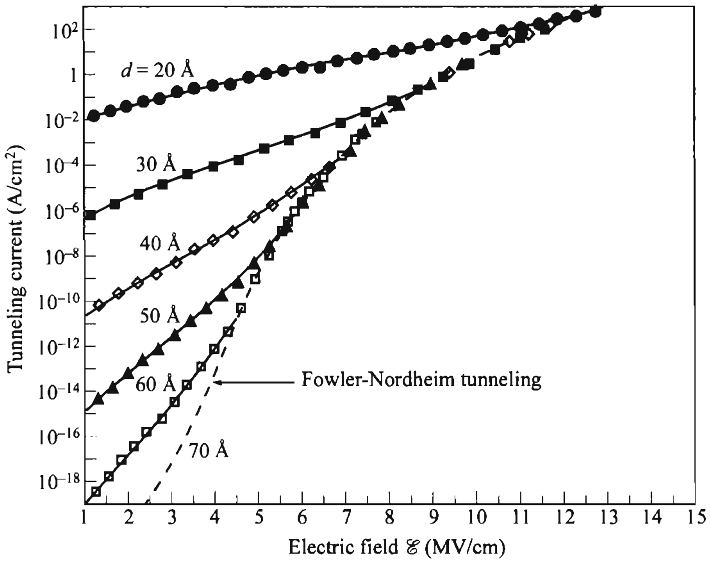


Fig. 18 Tunneling current vs. field for different oxide thicknesses. For thicker oxides Fowler-Nordheim tunneling dominates and is independent of thickness. Direct tunneling occurs in thinner oxides at low fields. (After Ref. 42.)

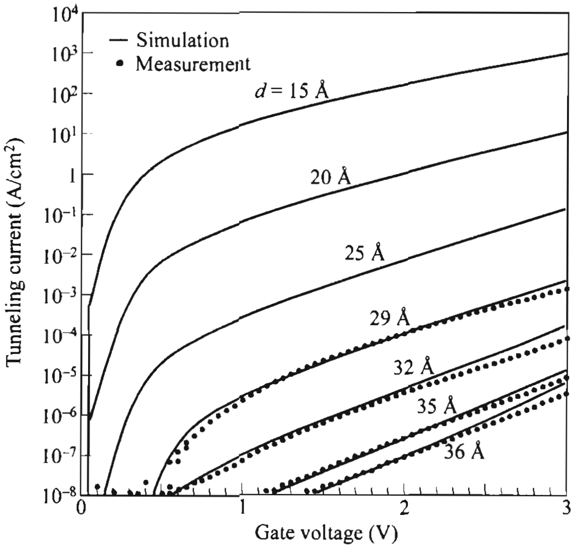


Fig. 19 Direct tunneling current taking quantum effects into account. (After Ref. 43.)

has a small depletion layer at the oxide interface which also increases the effective insulator thickness.

MIS Tunnel Diode. The tunneling current is given by an expression similar to that of Eq. 8. Using a WKB approximation and assuming the conservation of energy E and transverse momentum $k_\perp$, the tunneling current density along the x-direction between two conducting regions through a forbidden region can be written as[44]

$$J = \frac{q}{4\pi^2\hbar}\int\int T_t[F_1(E) - F_2(E)]dk_\perp^2 dE \tag{42}$$

where F_1 and F_2 are the Fermi distribution functions in the two conducting regions, and T_t is the tunneling probability. For the MIS diode under consideration, the constant energy surface in k-space for electrons in the semiconductor is, in general, considerably smaller than that in the metal. As a result, the tunneling of electrons from the semiconductor to the metal is always assumed to be allowable. If it is further assumed that the energy bands of the solids involved are parabolic with an isotropic electron mass m^*, Eq. 42 can be reduced to

$$J = \frac{m^* q}{2\pi^2\hbar^3}\int\int T_t dE_\perp dE \tag{43}$$

where $E_\perp$ and E are the transverse and total kinetic energy of electrons in the semiconductor. The limits of integration for $E_\perp$ are zero and E; the limits for E are simply the two Fermi levels. The tunneling probability for a rectangular barrier with an effective barrier height $q\phi_T$ and width d, Fig. 17c, can be obtained from Eq. 2:[45]

$$T_t \approx \exp\left(-\frac{2d\sqrt{2qm^*\phi_T}}{\hbar}\right)$$
$$\approx \exp(-\alpha_T d\sqrt{\phi_T}) \tag{44}$$

where $\alpha_T (= 2\sqrt{2qm^*}/\hbar)$ approaches unity if the effective mass in the insulator equals the free electron mass, and ϕ_T is in volts and d in Å.

The tunneling current can be evaluated by substituting Eq. 44 into Eq. 43 and integrating over the energy range, yielding[45,46]

$$J = A^*T^2\exp(-\alpha_T d\sqrt{q\phi_T})\exp\left(\frac{-q\phi_B}{kT}\right)\left[\exp\left(\frac{qV}{\eta kT}\right) - 1\right], \tag{45}$$

where $A^* = 4\pi m_t^* qk^2/h^3$ is the effective Richardson constant and ϕ_B the Schottky barrier height. Equation 45 is identical to the standard thermionic-emission equation for Schottky barriers except for the added term $\exp(-\alpha_T d\sqrt{q\phi_T})$, which is the tunneling probability. Here a constant of $[2(2m^*/\hbar^2)]^{1/2}$ is omitted which has the value of 1.01 $\text{eV}^{-1/2}\text{Å}^{-1}$. It is thus clear from Eq. 45 that for ϕ_T of the order of 1 V and $d > 50$ Å, the tunneling probability is about $\exp(-50) = 10^{-22}$, and the current is indeed negligibly small. As d and/or ϕ_T decrease, the current increases rapidly toward the thermionic-emission current level. Figure 20 shows the forward *I-V* characteristics of four Au-SiO_2-Si tunnel diodes of different insulator-layer thicknesses. For $d \approx 10$ Å, the current follows the standard Schottky-diode behavior with an ideality factor η close to 1. As the insulator thickness increases, the current decreases rapidly and the ideality factor begins to depart from unity. The expression for η has been presented in Section 3.3.6.

One of the most-important parameters for this MIS tunnel diode is the metal-insulator barrier height, which has a profound effect on the *I-V* characteristics.[47–48] Figure 21 shows the schematic energy-band diagrams at thermal equilibrium for MIS tunnel diodes on *p*-type substrates with two metal-to-insulator barrier heights. For the low-barrier case ($\phi_{mi} = 3.2$ V for the Al-SiO_2 system) the surface of the *p*-type silicon is inverted at equilibrium. Whereas in the high-barrier case ($\phi_{mi} = 4.2$ V for the

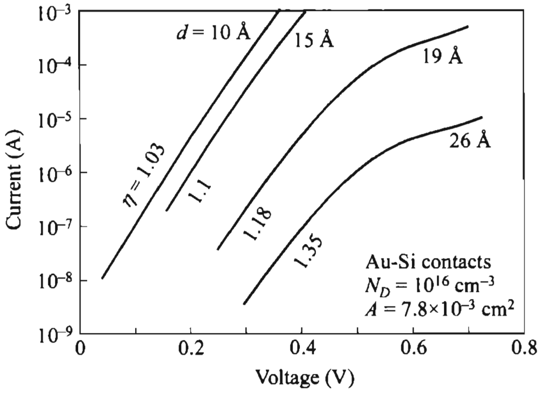


Fig. 20 Measured current-voltage characteristics of MIS tunnel diodes having different oxide thicknesses. (After Ref. 46.)

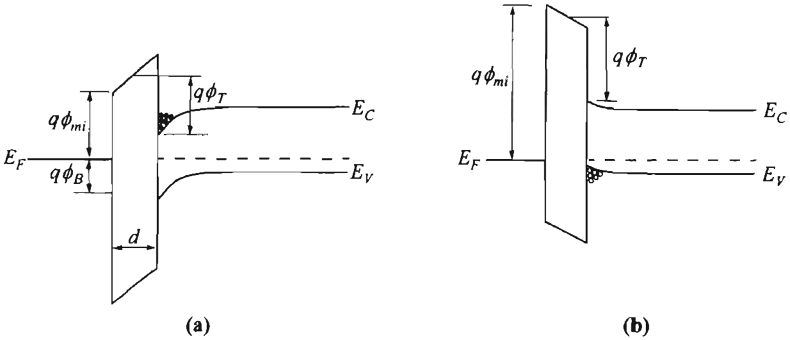


Fig. 21 Energy-band diagrams for MIS tunnel diodes on nondegenerate substrates (*p*-type) with (a) low metal-insulator barrier ϕ_{mi} and (b) high metal-insulator barrier. (After Ref. 47.)

Au-SiO_2 system) the surface is in accumulation of holes. Two main tunneling current components exist: J_{ct}, carriers from the conduction band to the metal, and J_{vt}, from the valence band to the metal. Both currents are given by expressions similar to Eq. 42.

Figure 22 shows the theoretical *I-V* curves for the two diodes. For the low-barrier case, Fig. 22a, under small forward and reverse biases, the dominant current is the minority-carrier (electron) current J_{ct}, due to the abundance of electrons. As the forward bias (positive voltage on semiconductor) increases, the current also increases monotonically. At a given bias, the current increases rapidly with decreasing insulator thickness. This is because the current is limited by the tunneling probability, Eq. 44,

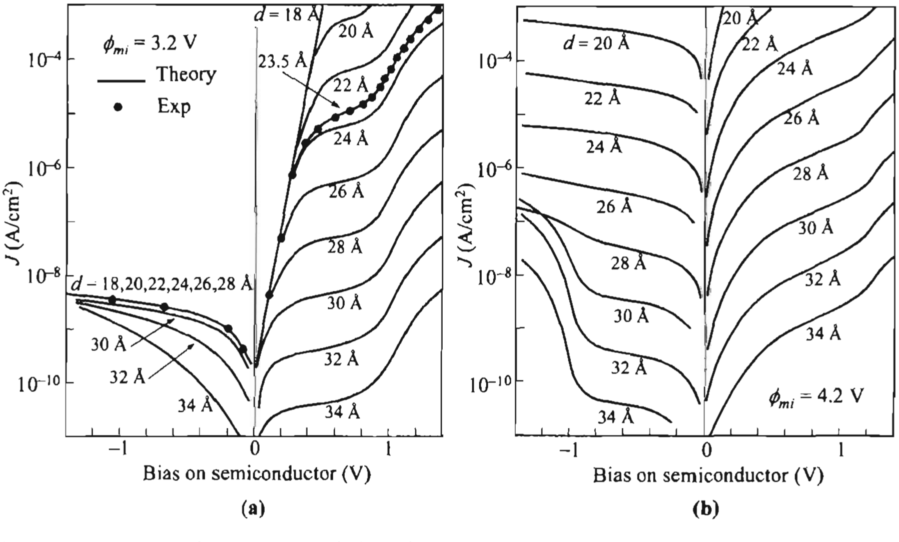


Fig. 22 Current-voltage characteristics of MIS tunnel diodes having: (a) low barrier ϕ_{mi}, (b) high barrier. T = 300 K. $N_A = 7\times10^{15}$ cm^{-3}. (After Ref. 47.)

which varies exponentially with the insulator thickness. At reverse bias, the current is virtually independent of the insulator thickness for $d < 30$ Å, because the current is now limited by the rate of supply of minority carriers (electrons) through the semiconductor, and the current is similar to the saturation current in a reverse-biased *p-n* junction. Figure 22a also shows the experimental result for $d = 23.5$ Å. Note that the current-voltage characteristics are similar to the rectification nature of a *p-n* junction.

For the high-barrier case, Fig. 22b, under forward bias, the dominant current is the majority-carrier (hole) tunneling current from the valence band to the metal, and the current increases exponentially with decreasing insulator thickness. Under reverse bias the current does not become independent of the insulator thickness as in Fig. 22a. Instead, the current increases rapidly with decreasing insulator thickness, because for majority-carrier transport the current is limited in both directions by the tunneling probability, not by the rate of carrier supply. So for high-barrier cases, the tunneling currents are higher, especially in reverse bias.

MIS Tunnel Diode On Degenerate Semiconductor. We discuss here that negative resistance can be observed from MIS tunnel diode on degenerately doped semiconductor. Figure 23 shows simplified band diagrams for MIS tunnel diodes with p^{++}- and n^{++}-semiconductor substrates, including interface traps. The band bending and image force on the semiconductor side and potential drops across the oxide layer at equilibrium are omitted for simplicity. Consider the p^{++}-type semiconductor first. Applying a positive voltage to the metal (Fig. 23b) causes electrons to tunnel from the valence band to the metal. The tunneling current in this bias polarity (Fig. 23b) increases monotonically with the increasing energy range between the Fermi levels and does not result in negative resistance; it further increases with the decreasing effective insulator barrier height ϕ_T.

Applying a small negative voltage to the metal (Fig. 23c) results in electrons tunneling from the metal to the unoccupied semiconductor valence band. According to Fig. 23d, for electrons tunneling from the metal to the unoccupied states of the valence band, an increase of the reverse voltage $-V$ implies an increase in the effective barrier height ϕ_T, thus resulting in a drop of current with increase of bias, i.e., negative resistance. Another current component results from electrons in the metal with higher energies tunneling simultaneously into the empty interface traps and momentarily recombining with holes in the valence band. Since the effective insulator barrier decreases with bias, this current component always has a positive differential resistance. Finally further increase of the bias results in a third, very fast-growing tunneling current component from the metal into the conduction band of the semiconductor (Fig. 23e).

Next consider the n^{++}-type semiconductors. As shown in Fig. 23f, the effective insulator barriers for the n^{++}-type are expected to be smaller than those of the p^{++}-type samples; hence, in general, for a given bias, there will be larger tunneling currents. For a negative bias on the metal, electrons tunnel from the metal into the empty states of the semiconductor conduction band, resulting in a large, rapidly increasing current (Fig. 23g). A small positive voltage on the metal leads to increasing electron tunneling from the conduction band of the semiconductor into the metal (Fig. 23h). If the

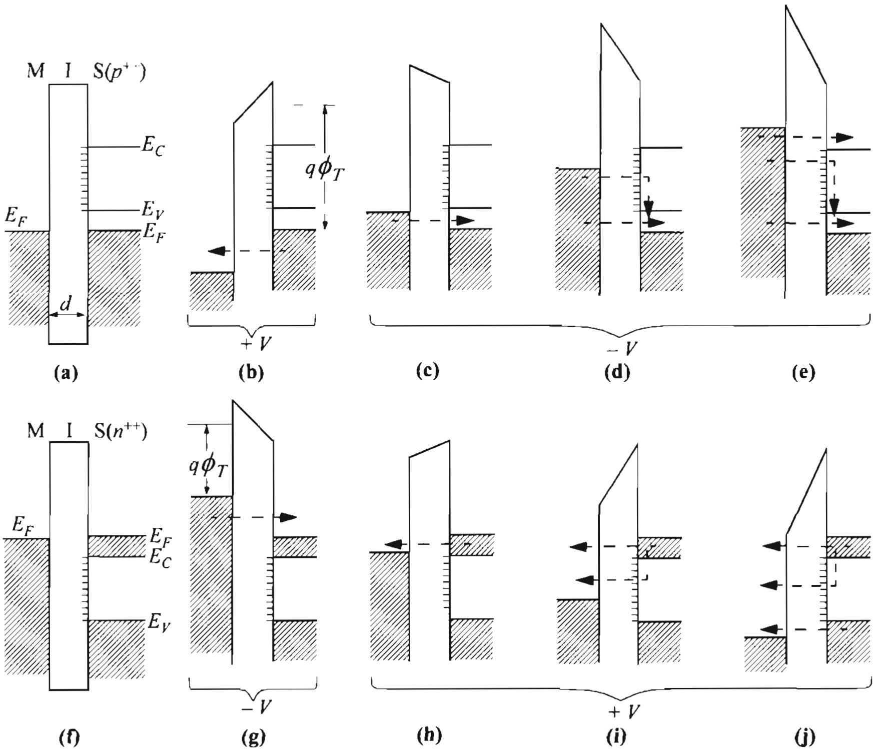


Fig. 23 Simplified band diagrams of MIS tunnel diodes on degenerate substrates, including interface traps. Top and bottom rows are for p^{++}- and n^{++}-substrates respectively. V is positive bias on the metal. (After Ref. 49.)

interface traps are filled with conduction electrons by recombination, a further increase in bias (Fig. 23i) gives rise to a second current component caused by the tunneling of electrons from the interface traps into the metal. This current component increases with increasing bias since the effective insulator barrier decreases. For a larger voltage (Fig. 23j) additional tunneling from the valence band to the metal is possible, but its influence on the total I-V characteristic is comparatively small because of the relatively high oxide barrier. Thus, the band structure of the semiconductor has a much smaller influence on the tunneling characteristics of the n^{++}-type compared to p^{++}-type structures. Note the interesting result that unlike p^{++}-substrate, there is no negative resistance observed on n^{++}-substrate.

The negative resistance on p^{++}-semiconductor has been obtained in MIS tunnel diodes of Al-Al_2O_3-SnTe.[50] The SnTe is a highly doped p-type with a concentration of 8×10^{20} cm^{-3}; the Al_2O_3 is about 5-nm thick. Figure 24 shows the measured current-voltage characteristics at three different temperatures, where the negative resistance occurs between 0.6 to 0.8 V. These results are in good agreement with theoretical prediction[44] based on Eq. 43.

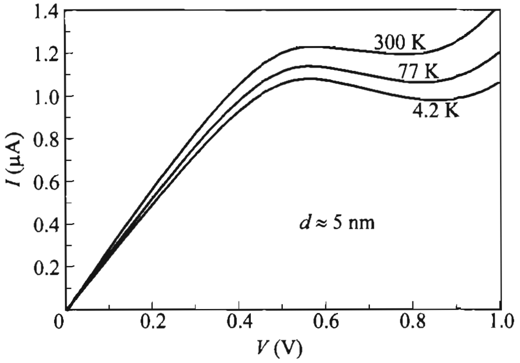
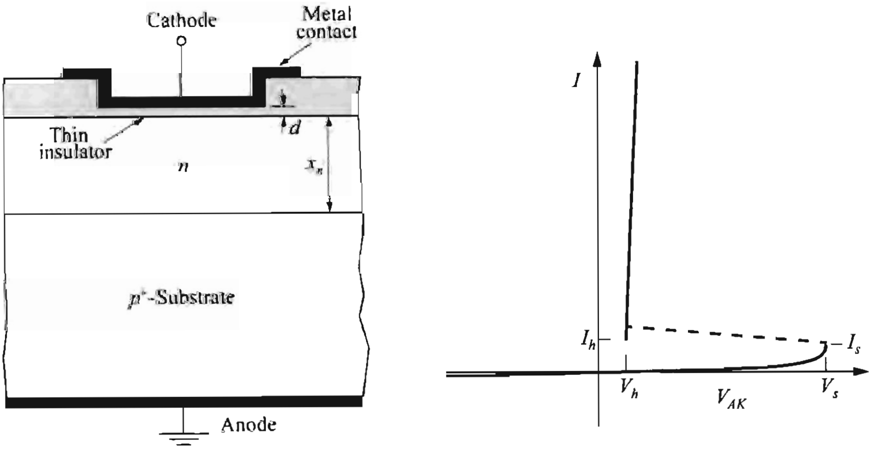


Fig. 24 MIS tunnel diode (Al-Al_2O_3-SnTe) I-V characteristics showing negative resistance. (After Ref. 50.)

8.3.3 MIS Switch Diode

The MIS switch (MISS) diode is a four-layer structure as shown in Fig. 25a. It is basically an MIS tunnel diode in series with a p-n junction. This diode was found to display a current-controlled negative resistance, Fig. 25b, similar to a Shockley diode (Chapter 11).[51] When a negative bias is applied to the top metal contact (or positive V_{AK}, p^+-region is assumed to be grounded), the I-V characteristic shows a high-impedance or *off-state*. At a sufficiently high voltage, the switching voltage V_s, the device suddenly switches to a low-voltage high-current *on-state*. The switching is initiated either by the extension of the surface depletion region to that of the p^+-n junction (punch-through), or by avalanche in the surface n-layer.[52] The initial device was built



(a)

(b)

Fig. 25 (a) MIS switch diode, a four-layer structure. (b) Current-voltage characteristics show current-controlled S-type negative resistance.

on a Si wafer and employed SiO_2 as the tunnel insulator. Later, similar behaviors were obtained from other insulators (e.g., Si_3N_4) and thick polycrystalline silicon.

The *I-V* characteristics shown in Fig. 25b can be explained qualitatively with the energy-band diagrams shown in Fig. 26. With negative anode-to-cathode voltage ($-V_{AK}$), the MIS tunnel diode is under forward bias and the *p-n* junction under reverse bias. The current is limited by generation within the depletion region (W_D) of the *p-n* junction, given by

$$J_g = \frac{qn_iW_D}{2\tau} \approx \frac{n_i}{\tau}\sqrt{\frac{q\varepsilon_s(|V_{AK}| + \psi_{bi})}{2N_D}}, \tag{46}$$

where τ is the minority-carrier lifetime and ψ_{bi} the built-in potential of the *p-n* junction. Under this bias condition, no switching occurs.

With positive V_{AK}, the MIS tunnel diode is under reverse bias and the *p-n* junction under forward bias (Fig. 26c). In the low-current off-state, current is dominated by generation in the surface depletion region, given by the same expression except that ψ_{bi} is replaced in this case by the barrier height ϕ_B at equilibrium. The thermally generated electrons approach the *p-n* junction and recombine with holes in the depletion region of the forward-biased *p-n* junction. This implies that the current through the *p-n* junction is dominantly recombination, rather than diffusion, due to the low current level passing through it. The electrons tunneling from metal to semiconductor is the reverse current of an MIS tunnel diode and it is small in the off-state. But this current will be shown later to be large and becomes the dominant current in the on-state.

The switching criterion of the MISS depends critically on the supply of holes toward the tunneling insulator. When this hole current is semiconductor-limited (generation) it is small. In this condition, the semiconductor surface is in deep depletion, and an inversion layer of holes at the surface is not formed. If an additional supply of holes from other sources is available, the tunneling current is not sufficient to drain the holes, so it becomes tunneling-limited and a hole inversion layer is formed. The collapse of the surface potential (surface band bending) increases the voltage across the insulator V_i and increases the J_{nt} in two respects. First, the barrier height ϕ_B is reduced, and second, ϕ_T is also reduced. The latter is equivalent to a higher electric field across the insulator. The large current passes through the *p-n* junction and the current mechanism in the *p-n* junction changes from recombination to diffusion. An electron current J_n can inject a much larger hole current since $N_A \gg N_D$, by a factor of $\approx 1/(1 - \gamma)$, where γ is the injection efficiency of the *p-n* junction (ratio of hole current to total current). The total hole current tunneling through the insulator becomes

$$J_{pt} = J_n\left(\frac{1}{1-\gamma}\right). \tag{47}$$

The MIS tunnel diode and the *p-n* junction pair creates regenerative feedback and results in negative differential resistance.

The regenerative feedback can also be viewed as a result of two current gains: a gain of electron current from hole current in an MIS tunnel diode, as originally pro-

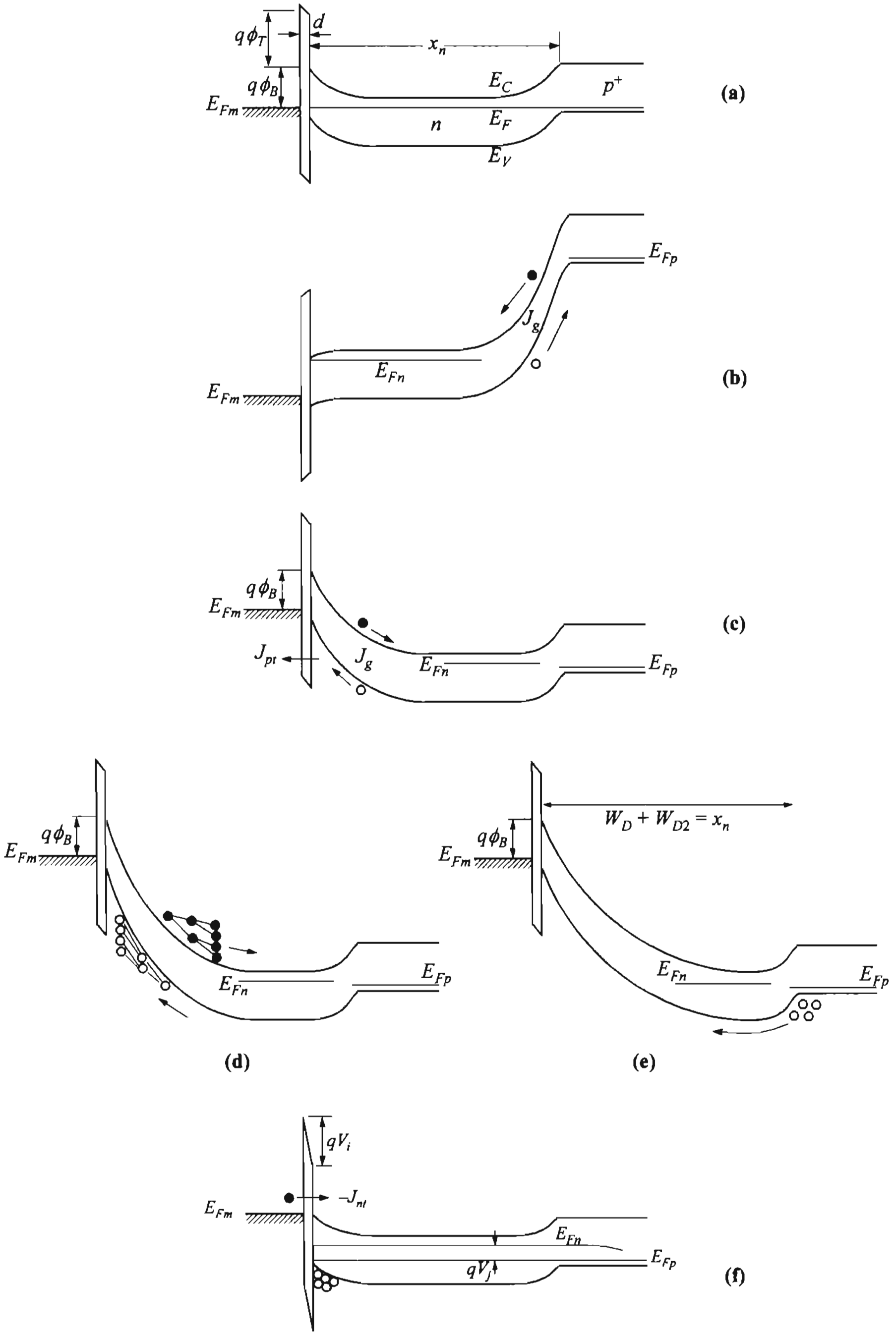

$q\phi_T$
d
x_n
$q\phi_B$
E_{Fm}
E_C
p^+
E_F
n
E_V
(a)
E_{Fp}
J_g
E_{Fn}
E_{Fm}
(b)
$q\phi_B$
E_{Fm}
J_{pt}
J_g
E_{Fn}
E_{Fp}
(c)
$q\phi_B$
E_{Fm}
E_{Fn}
E_{Fp}
(d)
$W_D + W_{D2} = x_n$
$q\phi_B$
E_{Fm}
E_{Fn}
E_{Fp}
(e)
qV_i
$-J_{nt}$
E_{Fm}
E_{Fn}
E_{Fp}
qV_j
(f)

Fig. 26 Energy-band diagrams of MISS under different biases. (a) Equilibrium. (b) Negative V_{AK}. (c) Positive V_{AK}, off-state. (d) Onset of avalanche multiplication. (e) Onset of punch-through. (f) High-current on-state.

posed,[53] and a gain of hole current from electron current in the p^+-n junction. To achieve current gain in an MIS tunnel diode, the precise insulator thickness is critical, and it has to be in the range of 20–40 Å for the case of silicon dioxide. Oxides thinner than 20 Å cannot confine holes at the surface to support an inversion layer and to decrease ϕ_B, and current is always semiconductor-limited. Oxides thicker than 50 Å do not allow deep depletion, and current is always tunneling-limited.

In practice, the current initiated by generation is not large enough to trigger switching. The two most-common additional sources are from punch-through and avalanche. In the punch-through condition shown in Fig. 26e, the depletion region of the MIS diode merges with that of the p-n junction. The potential barrier for holes is reduced and a large hole current is injected. The switching voltage in this punch-through mode is given by

$$V_s \approx \frac{qN_D(x_n - W_{D2})^2}{2\varepsilon_s} \tag{48}$$

where W_{D2} is the depletion region of the p-n junction.

Before the punch-through condition, if the electric field near the surface is high enough, avalanche multiplication occurs and also gives rise to a large hole current toward the surface (Fig. 26d). The switching voltage in this mode is similar to the avalanche breakdown voltage of a p-n junction. Avalanche-mode switching dominates in structures with high doping concentrations in the n-layer, usually higher than 10^{17} cm^{-3}.

The energy-band diagram in Fig. 26f shows an MISS after it is switched to the high-current on-state. Note that neither punch-through nor avalanche could be sustained after switching. The conduction-band edge at the surface is below E_{Fm} ($\phi_B = 0$), and J_{nt} controls the on-current. The holding voltage can be approximated by

$$V_h \approx V_i + V_j . \tag{49}$$

V_i is the voltage across the insulator, and is approximately equal to the original barrier height ϕ_B at equilibrium ($\approx$ 0.5–0.9 V). The forward bias on the p-n junction V_j is around 0.7 V, giving a holding voltage of $\approx$ 1.5 V.

Besides the aforementioned punch-through and avalanche, two other sources of hole current are also possible. One is by a third terminal contact, and another by optical generation. The three-terminal MISS is sometimes called an MIS thyristor. With either a minority- or majority-carrier injector, the function is the same—to increase the hole current flowing toward the insulator. While the minority-carrier injector injects holes directly, the majority-carrier injector controls the potential of the n-layer, and hole current is injected from the p^+-substrate. In either structure, with a positive gate current flowing into the device, a lower switching voltage results. Alternately when the MISS is exposed to a light source, J_p is generated optically and the switching voltage is reduced. For a fixed V_{AK}, light can induce turn-on and the device becomes a light-triggered switch.

As mentioned previously, the oxide thickness is a key parameter in the switching behavior. As shown in Fig. 27, for thicker oxides ($d \geq 50$ Å) the tunneling impedance is too high to meet the switching requirement. For very thin oxides ($d < 15$ Å) the

p^+-n junction can be turned fully on prior to the development of the deep depletion mode; thus the device displays a p-n junction characteristic. Switching behavior is observed only for the intermediate thicknesses (15 Å $< d <$ 40 Å).

Attractive features of the MIS switch diode include high switching speed (1 ns or less), and high sensitivity of the switching voltage V_S to light or current injection. The MISS can be applied in digital logic, and shift registers have been demonstrated. Other applications include memories such as SRAM, microwave generation when incorporated in a relaxation oscillator circuit, and as a light-triggered switch for alarm systems. The limitation of the MISS is its relatively high holding voltage and difficulty in reproducing a uniformly thin tunneling insulator.

8.3.4 MIM Tunnel Diode

A metal-insulator-metal (MIM) tunnel diode is a thin-film device in which the electrons from the first metal can tunnel into the insulator film and be collected by the second metal. It displays nonlinear I-V characteristics but negative resistance is not present. The nonlinear I-V nature is sometimes used for microwave detection as a mixer. Figures 28a and b show the basic energy-band diagrams of a MIM diode with similar metal electrodes. Since all of the voltage applied is dropped across the insulator, the tunneling current through the insulator is, from Eq. 42,

$$J = \frac{4\pi q m^*}{h^3}\int\int T_t[F(E) - F(E+qV)]dE_\perp dE\,. \qquad (50)$$

At 0 K, Equation 50 simplifies to[54]

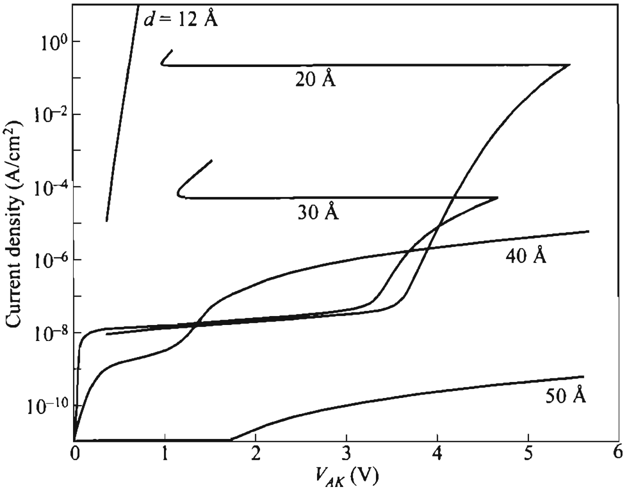


Fig. 27 Calculated I-V characteristic of MIS switch diodes for different values of oxide thickness. Device constants are x_n = 10 μm, N_D = 10^{14} cm^{-3}, and τ = 3.5×10^{-5} s. (After Ref. 52.)

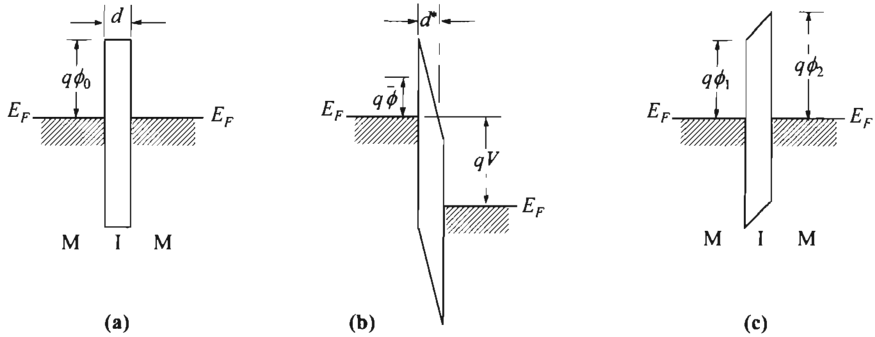


Fig. 28 Energy-band diagrams of MIM structures. (a) Symmetrical MIM under equilibrium. (b) Under bias, $V > \phi_0$. (c) Asymmetrical MIM.

$$J = J_0[\bar{\phi}\exp(-C\sqrt{\bar{\phi}}) - (\bar{\phi} + V)\exp(-C\sqrt{\bar{\phi} + V})] \tag{51}$$

where

$$J_0 \equiv \frac{q^2}{2\pi h d^{*2}}, \tag{52}$$

$$C \equiv \frac{4\pi d^* \sqrt{2m^* q}}{h}, \tag{53}$$

$\bar{\phi}$ is the average barrier height above the Fermi level, and d^* is the reduced effective barrier width. Equation 51 can be interpreted as a current density $J_0\bar{\phi}\exp(-C\sqrt{\bar{\phi}})$ flowing from electrode-1 to electrode-2 and another $J_0(\bar{\phi} + V)\exp(-C\sqrt{\bar{\phi} + V})$ flowing from electrode-2 to electrode-1.

We now apply Eq. 51 to the ideal symmetrical MIM structure. By ideal we mean that the temperature effect, the image-force effect, and the field-penetration effect in metal electrodes are neglected. For $0 \le V \le \phi_0$, $d^* = d$, and $\bar{\phi} = \phi_0 - V/2$, the current density is given by

$$J = J_0\left[\left(\phi_0 - \frac{V}{2}\right)\exp\left(-C\sqrt{\phi_0 - \frac{V}{2}}\right) - \left(\phi_0 + \frac{V}{2}\right)\exp\left(-C\sqrt{\phi_0 + \frac{V}{2}}\right)\right] \tag{54}$$

For larger voltage, $V > \phi_0$, we have $d^* = d\phi_0/V$ and $\bar{\phi} = \phi_0/2$. The current density is then

$$J = \frac{q^2\mathscr{E}^2}{4\pi h\phi_0}\left\{\exp\left(\frac{-\mathscr{E}_0}{\mathscr{E}}\right) - \left(1 + \frac{2V}{\phi_0}\right)\exp\left[\frac{-\mathscr{E}_0\sqrt{1 + (2V/\phi_0)}}{\mathscr{E}}\right]\right\} \tag{55}$$

where

$$\mathscr{E}_0 \equiv \frac{8\pi}{3h}\sqrt{2m^* q}\,\phi_0^{3/2}, \tag{56}$$

and $\mathscr{E} = V/d$ is the field in the insulator. For higher voltage such that $V > \phi_0$, the second term in Eq. 55 can be neglected, and we have the well-known Fowler-Nordheim tunneling equation (Eq. 41).

For an ideal asymmetrical MIM structure with different barrier heights ϕ_1 and ϕ_2 (Fig. 28c), in the low-voltage range $0 < V < \phi_1$, the quantities $d^* = d$ and $\bar{\phi} = (\phi_1 + \phi_2 - V)/2$ are independent of the polarities. Thus the J-V characteristics are also independent of the polarity. At higher voltages, $V > \phi_1$, the average barrier height $\bar{\phi}$ and the effective tunneling distance d^* become polarity-dependent. Therefore, the currents for different polarities are different.

The MIM tunnel diodes have been used to study the energy-momentum relation in the forbidden gap of large-bandgap semiconductors.[55,56] An MIM tunnel structure is formed using the single-crystal specimen, for example GaSe ($E_g = 2.0$ eV, $d < 10$ nm), sandwiched between two metal electrodes. Using one set of J-V curves, one can obtain the energy-momentum (E-k) relationship using Eqs. 42 and 50. Once the E-k relationship is obtained, one can calculate, using no adjustable parameters, the tunneling currents for all other thicknesses.

8.3.5 Hot-Electron Transistors

Over the years, many attempts have been made to invent or discover new solidstate devices capable of achieving better performance than bipolar transistors or MOSFETs. Among the most interesting candidates are the *hot-electron transistors* (HETs). In a hot-electron transistor, carriers injected from the emitter have high potential or kinetic energy in the base. Since a hot carrier has higher velocity, HETs are expected to have higher intrinsic speed, higher current, and higher transconductance. In this section, we discuss HETs based on tunnel emitter-base junctions. These devices are sometimes referred to as *tunneling hot-electron transfer amplifier* (THETA).

The first THETA was reported by Mead in 1960, using an MOMOM (metal-oxide-metal-oxide-metal) structure, sometimes called an MIMIM (metal-insulator-metal-insulator-metal) structure (Fig. 29a).[57,58] In this structure both the emitter and collector barriers are formed by oxides. The metal base had to be thin and was typically between 10 and 30 nm. The current gain of such a structure could be greatly improved by replacing the MOM collector junction by a metal-semiconductor junction (Fig. 29b),[59] resulting in an MOMS (metal-oxide-metal-semiconductor), or an MIMS (metal-insulator-metal-semiconductor) structure. This MIMS structure, however, has a lower maximum oscillation frequency than the bipolar transistor mainly because of its longer emitter charging time (caused by larger emitter capacitance) and smaller common-base current gain (caused by hot-electron scattering in the base region). Still another variation is to use a p-n junction in the collector (Fig. 29c).[60] In this MOp-n (or MIp-n) structure the semiconductor is the base material, as opposed to metal, and thus has less scattering in the base.

Since all the above structures use the same emitter injection mechanism of tunneling, they suffer from the same problem of low current gain and poor control of the barrier thickness. There has been a renewed interest in the THETA since Heiblum in

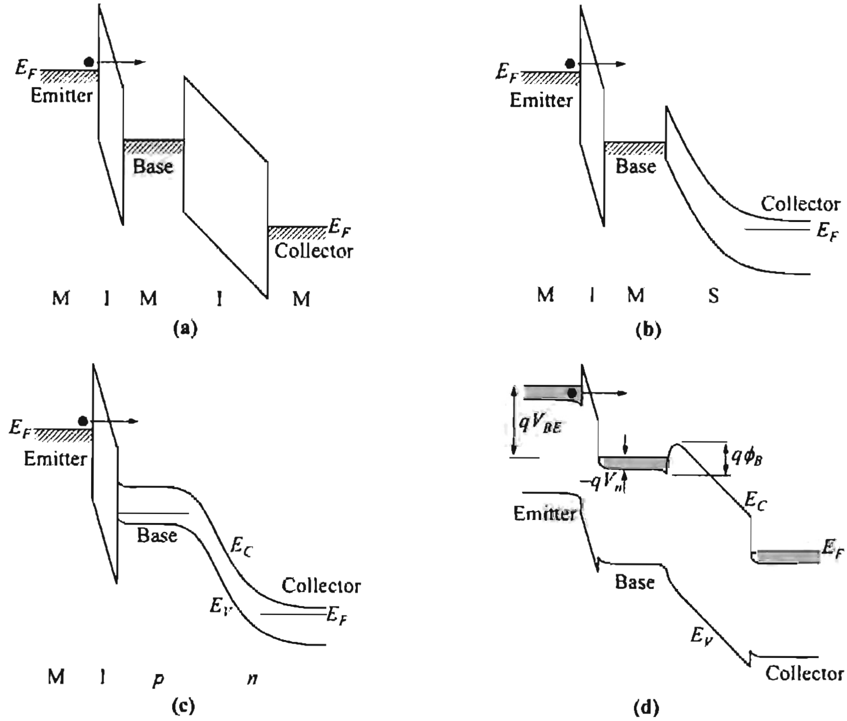


Fig. 29 Variations of tunnel-emitter hot-electron transistors and their energy-band diagrams under forward conditions. (a) MIMIM. (b) MIMS. (c) MI*p*-*n*. (d) Heterojunction THETA.

1981 suggested using a wide-energy-gap semiconductor as the tunneling barrier and a degenerately doped narrow-energy-gap semiconductor as the emitter, base, and collector.[61] This idea was especially timely after the rapid development of epitaxial techniques such as MBE and MOCVD in the 1970s. The first heterojunction THETA was reported in 1984,[62,63] followed by works reported in 1985.[64–66]

For the heterojunction structure (Fig. 29d), the AlGaAs/GaAs system is the most common, but other materials, such as InGaAs/InAlAs, InGaAs/InP, InAs/AlGaAsSb, and InGaAs/InAlGaAs, have been reported. The narrow-energy-gap materials for the emitter, base, and collector are typically heavily doped while the wide-energy-gap layers are undoped. The barrier thickness for the tunneling emitter is in the range of 7–50 nm, while the barrier layer for the collector is much thicker, ranging from 100 to 250 nm. The base width ranges from 10 to 100 nm. A thin base improves the transfer ratio but is harder to contact without shorting to the collector layer. The collector-base junction is often graded in composition to reduce quantum-mechanical reflection.

For the discussion of the working principle, the heterojunction THETA is assumed since it is of the greatest interest. Under normal operating conditions, the emitter is negatively biased (for the doping type shown) with respect to the base, and the collector is positively biased (Fig. 29d). Since the barrier created by the heterojunction is low, typically in the range 0.2–0.4 eV, it is necessary to operate the THETA at low temperatures to reduce the thermionic-emission current over the barrier. Electrons are injected from the emitter to the n^+-base, making the THETA a majority-carrier device. The emitter-base current is a tunneling current through the barrier, either by direct tunneling or by Fowler-Nordheim tunneling. The electrons injected at the base have a maximum kinetic energy (above E_C) of

$$E = q(V_{BE} - V_n) \tag{57}$$

(V_n is negative for a degenerate semiconductor). As electrons traverse the base, energy is lost through some scattering events. At the base-collector junction, carriers with energy above the barrier $q\phi_B$ will result in collector current, while the rest will contribute to undesirable base current.

The base transport factor α_T can be broken down into different components as

$$\begin{aligned} I_C &= \alpha_T I_E \\ &= \alpha_B \alpha_{BC} \alpha_C I_E \ . \end{aligned} \tag{58}$$

α_B, due to scattering in the base layer, is given by

$$\alpha_B = \exp\left(-\frac{W}{\lambda_m}\right) \tag{59}$$

where W and λ_m are the base width and its mean free path. Reported values for λ_m range from 70 to 280 nm. λ_m is also known to be dependent on the electron energy. When the energy is too high, λ_m starts to decrease. In the case of an MOM emitter, because the oxide barrier is much higher, a large V_{BE} is required to inject a specific current level. A higher V_{BE}, unfortunately, increases the electron energy and reduces λ_m. This is the factor that requires the oxide thickness in the MOM barrier to be small ($\approx$ 15 Å).[61] To improve α_B, the base thickness must be minimized, but this results in excessive base resistance. It has been suggested that an induced base[67] or modulation doping for the base layer be used so that it can be thin ($\approx$ 10 nm) and yet conductive. The second factor α_{BC} is due to quantum-mechanical reflection at the base-collector band-edge discontinuity. For an abrupt junction, it is given by[68]

$$\alpha_{BC} \approx 1 - \left[\frac{1 - \sqrt{1 - (q\phi_B/E)}}{1 + \sqrt{1 - (q\phi_B/E)}}\right]^2 . \tag{60}$$

Composition grading for the collector barrier would improve the reflection loss. α_C is the collector efficiency due to scattering in the wide-energy-gap material.

To have a high β (common-emitter current gain) value, α_T should be close to unity since

$$\beta \approx \frac{\alpha_T}{1 - \alpha_T} . \tag{61}$$

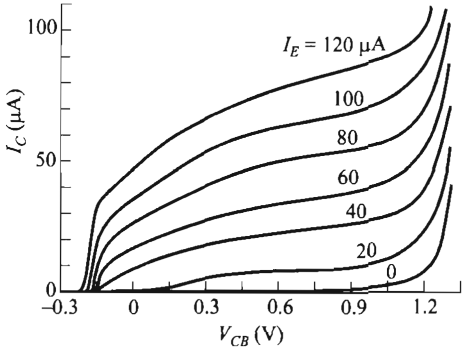
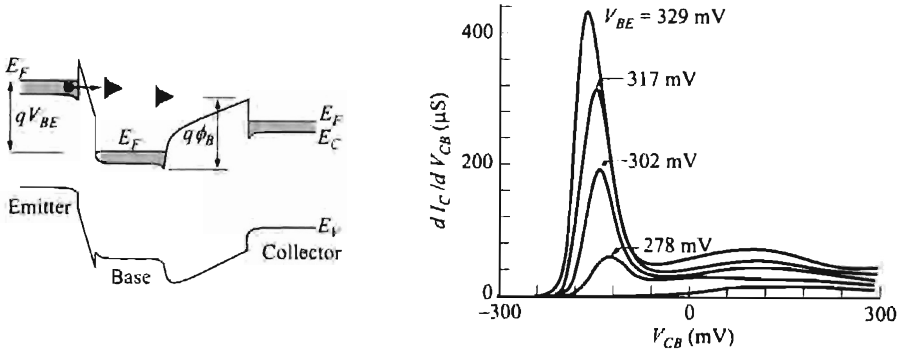


Fig. 30 Common-base output characteristics of a heterojunction THETA. (After Ref. 70.)

β values as high as 40 have been reported.[69] An example of the output characteristics of a THETA is shown in Fig. 30. The THETA offers a potential for high-speed operation due to ballistic transport through the base and the absence of minority-carrier storage. The requirement for lower-temperature operation is a concern that it may limit the application.

The THETA has been used as a research tool to study the properties of hot carriers. A specific function is a spectrometer to measure the energy spectrum of the tunneled hot electrons in the base. In this operation, the collector is biased positively with respect to the base to vary the effective collector barrier (Fig. 31a). When the incremental collector current is plotted against the effective collector barrier height, the energy spectrum of the hot electrons is obtained. It can be seen from Fig. 31b that for each V_{BE}, the energy (related to V_{CB}) at the peak of the distribution increases with V_{BE}.



(a) (b)

Fig. 31 (a) Energy-band diagram of THETA as a spectrometer. The collector voltage is negative with respect to the base to vary the effective collector barrier height. (b) Hot-electron energy spectrum. (After Ref. 70.)

8.4 RESONANT-TUNNELING DIODE

The negative differential resistance (NDR) of a resonant-tunneling diode (sometimes called a double-barrier diode) was predicted by Tsu and Esaki in 1973,[71] following their pioneering work on superlattices in the late 1960s and early 1970s. The structure and characteristics of this diode were first demonstrated by Chang et al. in 1974.[72] Following the much-improved results reported by numerous authors in the early 1980s, research interest escalated, partially due to the maturing MBE and MOCVD technologies. In 1985, room-temperature NDR in this structure was reported.[73–74]

A resonant-tunneling diode requires band-edge discontinuity at the conduction band or valence band to form a quantum well and thus necessitates heteroepitaxy. The most-popular material combination used is GaAs/AlGaAs (Fig. 32), followed by GaInAs/AlInAs. The middle quantum-well thickness is typically around 5 nm, and the barrier layers range from 1.5 to 5 nm. Symmetry of the barrier layers is not required, so their thicknesses can be different. The well and barrier layers are all undoped, and they are sandwiched between heavily doped, narrow-bandgap materials, which usually are the same as the well layer. Not shown in Fig. 32 are thin layers of undoped spacers (≈ 1.5 nm GaAs) adjacent to the barrier layers to ensure that dopants do not diffuse to the barrier layers.

Quantum mechanics prescribes that in a quantum well of width W, the conduction band (or valence band) is split into discrete subbands, and the bottom of each subband is given by

$$E_n - E_{Cw} = \frac{h^2 n^2}{8m^* W^2}, \qquad n = 1, 2, 3, \ldots \tag{62}$$

where E_{Cw} designates E_C in the well. Notice that this equation assumes infinite barrier height and can only serve to give a qualitative picture. In practice, the barrier (ΔE_C) lies in the range of 0.2–0.5 eV, giving quantized levels of ≈ 0.1 eV above E_C. Under

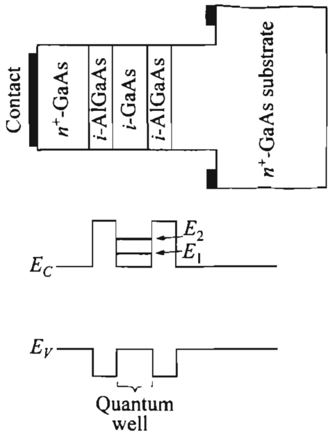


Fig. 32 Structure of resonant-tunneling diode using GaAs/AlGaAs heterostructure as an example. The energy-band diagram shows the formation of a quantum well and quantized levels.

bias conditions, carriers can tunnel from one electrode to the other via energy states within the well.

Resonant tunneling is a unique phenomenon when tunneling is through double barriers via quantized states.[75] Recall the case of tunneling through a single barrier, the tunneling probability is a monotonic increasing function with energy of the incoming particle. In resonant tunneling, the wavefunctions of the Schrödinger equation have to be solved simultaneously in the three regions—emitter, well, and collector. Because of the quantized states within the well, the tunneling probability exhibit peaks when the energy of the incoming particle coincides with one of the quantized levels, as shown in Fig. 33. In this coherent-tunneling picture, if the incoming energy does not coincide with any of the quantized levels, the tunneling probably is a product of the individual probability between the well and the emitter T_E, and that between the well and the collector T_C,

$$T(E) = T_E T_C. \tag{63}$$

However, when the incoming energy matches one of the quantized levels, the wavefunction builds up within the well similar to a Fabry-Perot resonator, and the transmission probability becomes[76]

$$T(E{=}E_n) = \frac{4T_E T_C}{(T_E + T_C)^2}. \tag{64}$$

For a symmetric structure, $T_E = T_C$, and $T = 1$. Away from this resonance, the value given by Eq. 63 quickly drops by many orders of magnitude, giving the shape shown in Fig. 33. The resonant-tunneling current is given by

$$J = \frac{q}{2\pi\hbar}\int N(E)T(E)dE \tag{65}$$

where the number of available electrons for tunneling (per unit area) from the emitter can be shown to be[75]

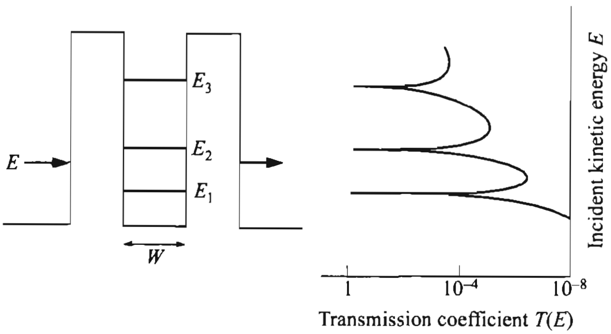


Fig. 33 Transmission coefficient of electron with energy E through a double barrier via coherent resonant tunneling. Transmission peaks occur when E aligns with E_n. (After Ref. 75.)

$$N(E) = \frac{kTm^*}{\pi\hbar^2}\ln\left[1+\exp\left(\frac{E_F-E}{kT}\right)\right]. \tag{66}$$

In a resonant-tunneling diode, the variable energy of the incoming electrons is provided by external bias such that the emitter energy is raised with respect to the well and the collector. The incoming energy distribution of tunneling electrons integrated over the sharp resonant-tunneling peaks of Fig. 33 would seem to predict sharp current peaks and very high peak-to-valley ratios, which are not observed in real devices, even at low temperatures. The reason for this is two-fold. First, the resonant transmission peaks are exponentially narrow, on the order of $\Delta E = \hbar/\tau$, where τ is the lifetime of an electron in the subband E_n with respect to tunneling out and ΔE is the broadening of the energy level E_n.[75] Additionally there exist nonideal effects such as impurity scattering, inelastic phonon scattering, phonon-assisted tunneling, and thermionic emission over the barrier. These effects lead to much larger valley current that diminishes the peak-to-valley ratio. As it turns out, a model of sequential tunneling, rather than coherent tunneling, can explain the experimental data quite well.[77] In the sequential-tunneling picture, the tunneling from emitter into the well, and that from the well to the collector, can be treated as uncorrelated events. This simpler picture can give better insight into the observed experimental data and is used below.

The I-V characteristics of a resonant-tunneling diode are shown qualitatively in Fig. 34. One notes not only the negative resistance, but also that it can be repeated, with multiple current peaks and valleys. This feature is not present in a conventional p^+-n^+ tunnel diode. The energy-band diagrams under biases that correspond to the different regions of the I-V curve are shown in Fig. 35. The peak current corresponds to the bias condition that E_C of the emitter electrode is in line with each quantized level. We next explain the origin of such negative resistance.

In the model of sequential tunneling, tunneling of carriers out of the well to the collector is much less constrained, and tunneling of carriers from the emitter into the well is the determining mechanism for the current flow. This requires available empty states at the same energy level (conservation of energy) and with the same lateral momentum (conservation of momentum) of the available electrons in the emitter as within the well. Since the parallel (to tunneling direction) momentum k_x in a quantum well is quantized which gives rise to quantized level E_n (i.e., $\hbar^2k_x^2/2m^* = h^2n^2/8m^*W^2$), the energy of carriers in each subband is a function of the lateral momentum $k_\perp$ only, given by

$$E_w = E_n + \frac{\hbar^2k_\perp^2}{2m^*}. \tag{67}$$

From Eq. 67 it should be noted that the energy of carriers is quantized only for the bottom of the subband, but the energy above E_n is continuous. The free-electron energy in the emitting electrode is, on the other hand, given by

$$E = E_C + \frac{\hbar^2k^2}{2m^*} = E_C + \frac{\hbar^2k_x^2}{2m^*} + \frac{\hbar^2k_\perp^2}{2m^*}. \tag{68}$$

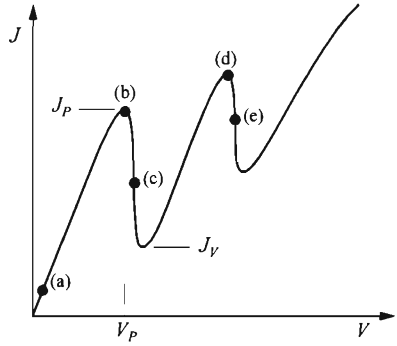


Fig. 34 *I-V* characteristics of resonant-tunneling diode with multiple current peaks and valleys at low but finite temperature. The labels (a)–(e) correspond to the energy-band diagrams shown in Fig. 35.

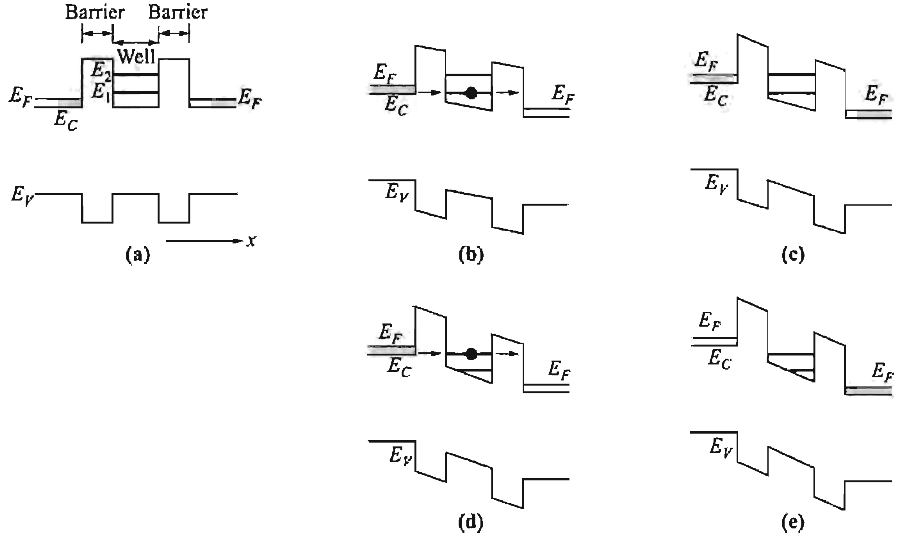


Fig. 35 Energy-band diagrams of resonant-tunneling diode under various biases. (a) Near zero bias. (b) Resonant tunneling through E_1. (c) E_1 below E_C. First region of NDR. (d) Resonant tunneling through E_2. (e) E_2 below E_C. Second region of NDR. Their corresponding electrical characteristics are shown in Fig. 34.

So electrons in the emitter with energy given by Eq. 68 will tunnel into the energy level given in Eq. 67. This concept is depicted in Fig. 36.

We first examine Region-a of the *I-V* curve in Fig. 34, where current increases with bias. Figure 36a shows that if E_1 is above E_F, there is little availability of electrons for tunneling. As the bias is increased, E_1 is pulled below E_F and toward E_C of the emitter, and tunneling current starts to increase with bias.

The decrease of current with bias in Region-c of Fig. 34 is less trivial. Conservation of lateral momentum requires that the last terms of Eqs. 67 and 68 are equal. This, along with the conservation of energy, results in the requirement

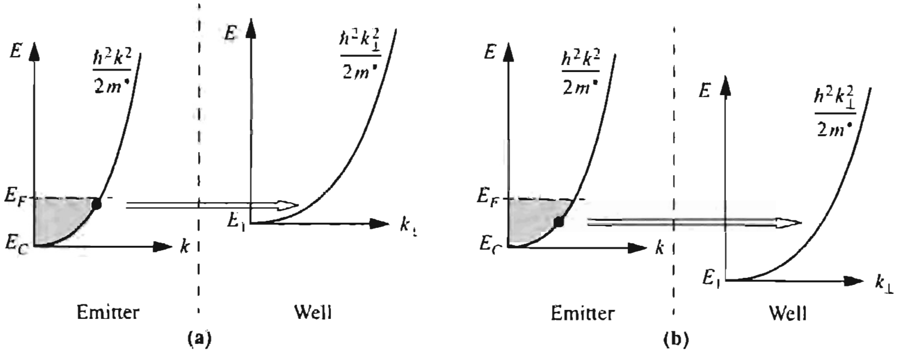

E
$\frac{\hbar^2 k^2}{2m^*}$
E_F
E_C
k
Emitter
E
$\frac{\hbar^2 k_\perp^2}{2m^*}$
E_1
$k_\perp$
Well
(a)
E
$\frac{\hbar^2 k^2}{2m^*}$
E_F
E_C
k
Emitter
E
$\frac{\hbar^2 k_\perp^2}{2m^*}$
E_1
$k_\perp$
Well
(b)

Fig. 36 Tunneling of electrons from emitter electrode into the well. Note the difference of k vs. $k_\perp$ in abscissas. (a) E_1 is higher than E_C and lower than E_F. Resonant tunneling starts to occur. (b) E_1 is lower than E_C. Tunneling probability drops significantly.

$$E_C + \frac{\hbar^2 k_x^2}{2m^*} = E_n \quad . \tag{69}$$

The energy equation implies that as long as the emitter E_C is above E_n, resonant tunneling is possible. With the help of Fig. 36b, it can be explained that this is not the case when momentum is taken into account. From this figure, it is seen that $k_\perp$ at the well becomes large. So for the emitter, since

$$k^2 = k_x^2 + k_\perp^2 , \tag{70}$$

even with $k_x = 0$, the minimum value of k is $k_\perp$. At low temperature, electrons are within the Fermi sphere of finite momentum. For $k_\perp$ outside the Fermi sphere, there is no electron available for tunneling. So the tunneling event in Fig. 36b is prohibited.

From the above discussion, for maximum tunneling current, E_n should line up between E_F and E_C of the emitter. But at low temperatures, E_n should line up with E_C, as shown in the bias conditions of Figs. 35b and d. With higher bias, the emitter E_C is above E_n and tunneling current is significantly reduced, resulting in NDR. The peak voltage occurs approximately at a voltage

$$V_P \approx \frac{2(E_n - E_C)}{q} \tag{71}$$

for a symmetrical junction, because half of the bias is developed across each barrier. In practical devices, V_P is larger than that given by Eq. 71. (Note that E_C is for the emitter in case it is different from the well.) The electric field can penetrate into both the emitter and collector regions and results in some voltage drop. Second, there is also a voltage drop across each undoped spacer layer. Another effect is caused by the finite charge accumulated within the well under bias. This charge sheet creates an inequity of electric field across the two barriers, and extra voltage is required to shift the relative energy between the emitter and the well.

The ratio of local peak current (J_P) to valley current (J_V) is a critical measure of the NDR. The current is modified from Eq. 65 to be

$$J = \frac{qN(V)T_E(V)\Delta E}{2\pi\hbar} \approx \frac{qN(V)T_E(V)}{2\pi\tau} \tag{72}$$

which can be maximized by using material of lighter effective mass. In this respect, the material combination of GaInAs/AlInAs is advantageous over GaAs/AlGaAs. A maximum peak-current density in the mid-10^5 A/cm^2 range has been observed and is quite temperature independent since it is a tunneling current. The nonzero valley current is due mainly to thermionic emission over the barriers, and it has a large temperature dependence (smaller J_V with lower temperature). Another small but conceivable contribution is due to tunneling of electrons to higher quantized levels. Even though the number of electrons available for tunneling at energy higher than E_F is very small, there is a thermal distribution tail and this number is not zero, especially when the quantized levels are close together.

As an example, the characteristics shown in Fig. 34 have two regions of NDR for each voltage polarity. In practice, the second current peak is rarely observed, due to the small signal in a large background of thermionic-emission current. The illustration nevertheless brings out the potential advantage over a tunnel diode that is limited to only one region of NDR. This feature of multiple current peaks is especially important as a functional device, which can perform more complex functions with a single device whereas conventional design would require many more components.

Because tunneling is inherently a very fast phenomenon that is not transit-time limited, the resonant-tunneling diode is considered among the fastest devices. Furthermore, it does not suffer from minority-charge storage. It has been demonstrated as an oscillator that can generate 700-GHz signals.[78] Maximum operational oscillation frequency has been projected to be over 1 THz. On the other hand, it is more difficult to use tunneling to supply high current, and the output power of an oscillator is limited. The resonant-tunneling diode has also been used in fast pulse-forming circuits and trigger circuits.[79] The unique feature of multiple current peaks can result in efficient functional devices. Examples are multivalue logic and memory.[80] The resonant-tunneling diode also serves as the building block for other three-terminal devices, such as the resonant-tunneling bipolar transistor and the resonant-tunneling hot-electron transistor.[81] It has also been incorporated in structures to study hot-electron spectroscopy.[82]

REFERENCES

1. K. K. Thornber, T. C. McGill, and C. A. Mead, "The Tunneling Time of an Electron," *J. Appl. Phys.*, **38**, 2384 (1967).

2. J. Niu, S. Y. Chung, A. T. Rice, P. R. Berger, R. Yu, P. E. Thompson, and R. Lake, "151 kA/cm^2 Peak Current Densities in Si/SiGe Resonant Interband Tunneling Diodes for High-Power Mixed-Signal Applications," *Appl. Phys. Lett.*, **83**, 3308 (2003).
3. P. Chahal, F. Morris, and G. Frazier, "50 GHz Resonant Tunneling Diode Relaxation Oscillator," *2004 Dev. Res. Conf. Digest*, p. 241.
4. Q. Liu and A. Seabaugh, "Design Approach Using Tunnel Diodes for Lowering Power in Differential Amplifiers," *IEEE Trans. Circ. Sys. –II: Express Briefs*, **52**, 572 (2005).
5. L. Esaki, "New Phenomenon in Narrow Germanium *p-n* Junctions," *Phys. Rev.*, **109**, 603 (1958).
6. L. Esaki, "Long Journey into Tunneling," *Proc. IEEE*, **62**, 825 (1974).
7. L. Esaki, "Discovery of the Tunnel Diode," *IEEE Trans. Electron Dev.*, **ED-23**, 644 (1976).
8. N. Holonyak, Jr. and I. A. Lesk, "Gallium Arsenide Tunnel Diodes," *Proc. IRE*, **48**, 1405 (1960).
9. R. L. Batdorf, G. C. Dacey, R. L. Wallace, and D. J. Walsh, "Esaki Diode in InSb," *J. Appl. Phys.*, **31**, 613 (1960).
10. A. G. Chynoweth, W. L. Feldmann, and R. A. Logan, "Excess Tunnel Current in Silicon Esaki Junctions," *Phys. Rev.*, **121**, 684 (1961).
11. H. P. Kleinknecht, "Indium Arsenide Tunnel Diodes," *Solid-State Electron.*, **2**, 133 (1961).
12. W. N. Carr, "Reversible Degradation Effects in GaSb Tunnel Diodes," *Solid-State Electron.*, **5**, 261 (1962).
13. C. A. Burrus, "Indium Phosphide Esaki Diodes," *Solid-State Electron.*, **5**, 357 (1962).
14. R. N. Hall, "Tunnel Diodes," *IRE Trans. Electron Devices*, **ED-7**, 1 (1960).
15. W. B. Joyce and R. W. Dixon, "Analytic Approximations for the Fermi Energy of an Ideal Fermi Gas," *Appl. Phys. Lett.*, **31**, 354 (1977).
16. J. V. Morgan and E. O. Kane, "Observation of Direct Tunneling in Germanium," *Phys. Rev. Lett.*, **3**, 466 (1959).
17. L. D. Landau and E. M. Lifshitz, *Quantum Mechanics*, Addison-Wesley, Reading, Mass., 1958, p. 174.
18. E. O. Kane, "Theory of Tunneling," *J. Appl. Phys.*, **32**, 83 (1961); "Tunneling in InSb," *J. Phys. Chem. Solids*, **12**, 181 (1960).
19. P. N. Butcher, K. F. Hulme, and J. R. Morgan, "Dependence of Peak Current Density on Acceptor Concentration in Germanium Tunnel Diodes," *Solid-State Electron.*, **5**, 358 (1962).
20. T. A. Demassa and D. P. Knott, "The Prediction of Tunnel Diode Voltage-Current Characteristics," *Solid-State Electron.*, **13**, 131 (1970).
21. D. Meyerhofer, G. A. Brown, and H. S. Sommers, Jr., "Degenerate Germanium I, Tunnel, Excess, and Thermal Current in Tunnel Diodes," *Phys. Rev.*, **126**, 1329 (1962).
22. L. V. Keldysh, "Behavior of Non-Metallic Crystals in Strong Electric Fields," *Sov. J. Exp. Theor. Phys.*, **6**, 763 (1958).
23. D. K. Roy, "On the Prediction of Tunnel Diode *I-V* Characteristics," *Solid-State Electron.*, **14**, 520 (1971).
24. W. N. Carr, "Reversible Degradation Effects in GaSb Tunnel Diodes," *Solid-State Electron.*, **5**, 261 (1962).

25. S. Ahmed, M. R. Melloch, E. S. Harmon, D. T. McInturff, and J. M. Woodall, "Use of Non-stoichiometry to Form GaAs Tunnel Junctions," *Appl. Phys. Lett.*, **71**, 3667 (1997).
26. V. M. Franks, K. F. Hulme, and J. R. Morgan, "An Alloy Process for Making High Current Density Silicon Tunnel Diode Junction," *Solid-State Electron.*, **8**, 343 (1965).
27. R. M. Minton and R. Glicksman, "Theoretical and Experimental Analysis of Germanium Tunnel Diode Characteristics," *Solid-State Electron.*, **7**, 491 (1964).
28. R. A. Logan, W. M. Augustyniak, and J. F. Gilber, "Electron Bombardment Damage in Silicon Esaki Diodes," *J. Appl. Phys.*, **32**, 1201 (1961).
29. W. Bernard, W. Rindner, and H. Roth, "Anisotropic Stress Effect on the Excess Current in Tunnel Diodes," *J. Appl. Phys.*, **35**, 1860 (1964).
30. V. V. Galavanov and A. Z. Panakhov, "Influence of Hydrostatic Pressure on the Tunnel Current in GaSb Diodes," *Sov. Phys. Semicond.*, **6**, 1924 (1973).
31. R. E. Davis and G. Gibbons, "Design Principles and Construction of Planar Ge Esaki Diodes," *Solid-State Electron.*, **10**, 461 (1967).
32. L. Esaki and Y. Miyahara, "A New Device Using the Tunneling Process in Narrow *p-n* Junctions," *Solid-State Electron.*, **1**, 13 (1960).
33. R. N. Hall, J. H. Racette, and H. Ehrenreich, "Direct Observation of Polarons and Phonons During Tunneling in Group 3-5 Semiconductor Junctions," *Phys. Rev. Lett.*, **4**, 456 (1960).
34. A. G. Chynoweth, R. A. Logan, and D. E. Thomas, "Phonon-Assisted Tunneling in Silicon and Germanium Esaki Junctions," *Phys. Rev.*, **125**, 877 (1962).
35. J. B. Hopkins, "Microwave Backward Diodes in InAs," *Solid-State Electron.*, **13**, 697 (1970).
36. A. B. Bhattacharyya and S. L. Sarnot, "Switching Time Analysis of Backward Diodes," *Proc. IEEE*, **58**, 513 (1970).
37. S. T. Eng, "Low-Noise Properties of Microwave Backward Diodes," *IRE Trans. Microwave Theory Tech.*, **MTT-8**, 419 (1961).
38. H. C. Torrey and C. A. Whitmer, *Crystal Rectifiers*, McGraw-Hill, New York, 1948. Ch. 8.
39. S. M. Sze and R. M. Ryder, "The Nonlinearity of the Reverse Current-Voltage Characteristics of a *p-n* Junction near Avalanche Breakdown," *Bell Syst. Tech. J.*, **46**, 1135 (1967).
40. J. Karlovsky, "The Curvature Coefficient of Germanium Tunnel and Backward Diodes," *Solid-State Electron.*, **10**, 1109 (1967).
41. M. Lenzlinger and E. H. Snow, "Fowler-Nordheim Tunneling into Thermally Grown SiO_2," *J. Appl. Phys.*, **40**, 278 (1969).
42. W. K. Shih, E. X. Wang, S. Jallepalli, F. Leon, C. M. Maziar, and A. F. Tasch, Jr., "Modeling Gate Leakage Current in nMOS Structures due to Tunneling Through an Ultra-Thin Oxide," *Solid-State Electron.*, **42**, 997 (1998).
43. S. H. Lo, D. A Buchanan, Y. Taur, and W. Wang, "Quantum-Mechanical Modeling of Electron Tunneling Current from the Inversion Layer of Ultra-Thin-Oxide nMOSFET's," *IEEE Electron Dev. Lett.*, **EDL-18**, 209 (1997).
44. L. L. Chang, P. J. Stiles, and L. Esaki, "Electron Tunneling between a Metal and a Semiconductor: Characteristics of Al-Al_2O_3-SnTe and -GeTe Junctions," *J. Appl. Phys.*, **38**, 4440 (1967).
45. V. Kumar and W. E. Dahlke, "Characteristics of Cr-SiO_2-*n*Si Tunnel Diodes," *Solid-State Electron.*, **20**, 143 (1977).

46. H. C. Card and E. H. Rhoderick, "Studies of Tunnel MOS Diodes I. Interface Effects in Silicon Schottky Diodes," *J. Phys. D: Appl. Phys.*, **4**, 1589 (1971).
47. M. A. Green, F. D. King, and J. Shewchun, "Minority Carrier MIS Tunnel Diodes and Their Application to Electron and Photovoltaic Energy Conversion: I. Theory," *Solid-State Electron.*, 17, 551 (1974). "II. Experiment," *Solid-State Electron.*, **17**, 563 (1974).
48. V. A. K. Temple, M. A. Green, and J. Shewchun, "Equilibrium-to-Nonequilibrium Transition in MOS Tunnel Diodes," *J. Appl. Phys.*, **45**, 4934 (1974).
49. W. E. Dahlke and S. M. Sze, "Tunneling in Metal-Oxide-Silicon Structures," *Solid-State Electron.*, **10**, 865 (1967).
50. L. Esaki and P. J. Stiles, "New Type of Negative Resistance in Barrier Tunneling," *Phys. Rev. Lett.*, **16**, 1108 (1966).
51. T. Yamamota and M. Morimoto, "Thin-MIS-Structure Si Negative Resistance Diode," *Appl. Phys. Lett.*, **20**, 269 (1972).
52. S. E.-D. Habib and J. G. Simmons, "Theory of Switching in *p-n* Insulator (Tunnel)-Metal Devices," *Solid-State Electron.*, **22**, 181 (1979).
53. M. A. Green and J. Shewchun, "Current Multiplication in Metal-Insulator-Semiconductor (MIS) Tunnel Diodes," *Solid-State Electron.*, **17**, 349 (1974).
54. J. G. Simmons, "Generalized Formula for the Electric Tunnel Effect between Similar Electrodes Separated by a Thin Insulating Film," *J. Appl. Phys.*, **34**, 1793 (1963).
55. S. Kurtin, T. C. McGill, and C. A. Mead, "Tunneling Currents and *E-k* Relation," *Phys. Rev. Lett.*, **25**, 756 (1970).
56. S. Kurtin, T. C. McGill, and C. A. Mead, "Direct Interelectrode Tunneling in GaSe," *Phys. Rev.*, **B3**, 3368 (1971).
57. C. A. Mead, "Tunnel-Emission Amplifiers," *Proc. IRE*, **48**, 359 (1960).
58. C. A. Mead, "Operation of Tunnel-Emission Devices," *J. Appl. Phys.*, **32**, 646 (1961).
59. J. P. Spratt, R. F. Schwartz, and W. M. Kane, "Hot Electrons in Metal Films: Injection and Collection," *Phys. Rev. Lett.*, **6**, 341 (1961).
60. H. Kisaki, "Tunnel Transistor," *Proc. IEEE*, **61**, 1053 (1973).
61. M. Heiblum, "Tunneling Hot Electron Transfer Amplifiers (THETA): Amplifiers Operating up to the Infrared," *Solid-State Electron.*, **24**, 343 (1981).
62. N. Yokoyama, K. Imamura, T. Ohshima, H. Nishi, S. Muto, K. Kondo, and S. Hiyamizu, "Tunneling Hot Electron Transistor using GaAs/AlGaAs Heterojunctions," *Jpn. J. Appl. Phys.*, **23**, L311 (1984).
63. N. Yokoyama, K. Imamura, T. Ohshima, H. Nishi, S. Muto, K. Kondo, and S. Hiyamizu, "Characteristics of Double Heterojunction GaAs/AlGaAs Hot Electron Transistors," *Tech. Dig. IEEE IEDM*, 532 (1984).
64. M. Heiblum, D. C. Thomas, C. M. Knoedler, and M. I. Nathan, "Tunneling Hot-Electron Transfer Amplifier: A Hot-Electron GaAs Device with Current Gain," *Appl. Phys. Lett.*, **47**, 1105 (1985).
65. M. Heiblum and M. V. Fischetti, "Ballistic Electron Transport in Hot Electron Transistors," in F. Capasso, Ed., *Physics of quantum electron devices*, Springer-Verlag, New York, 1990.
66. I. Hase, H. Kawai, S. Imanaga, K. Kaneko, and N. Watanabe, "MOCVD-Grown AlGaAs/GaAs Hot-Electron Transistor with a Base Width of 30 nm," *Electron. Lett.*, **21**, 757 (1985).

67. S. Luryi, "Induced Base Transistor," *Physica*, **134B**, 466 (1985).

68. S. Luryi, "Hot-Electron Injection and Resonant-Tunneling Heterojunction Devices," in F. Capasso and G. Margaritondo, Eds., *Heterojunction Band Discontinuities: Physics and Device Applications*, Elsevier Science, New York, 1987.

69. K. Seo, M. Heiblum, C. M. Knoedler, J. E. Oh, J. Pamulapati, and P. Bhattacharya, "High-Gain Pseudomorphic InGaAs Base Ballistic Hot-Electron Device," *IEEE Electron Dev. Lett.*, **EDL-10**, 73 (1989).

70. M. Heiblum, M. I. Nathan, D. C. Thomas, and C. M. Knoedler, "Direct Observation of Ballistic Transport in GaAs," *Phys. Rev. Lett.*, **55**, 2200 (1985).

71. R. Tsu and L. Esaki, "Tunneling in a Finite Superlattice," *Appl. Phys. Lett.*, **22**, 562 (1973).

72. L. L. Chang, L. Esaki, and R. Tsu, "Resonant Tunneling in Semiconductor Double Barriers," *Appl. Phys. Lett.*, **24**, 593 (1974).

73. T. J. Shewchuk, P. C. Chapin, and P. D. Coleman, "Resonant Tunneling Oscillations in a GaAs-$Al_xGa_{1-x}As$ Heterostructure at Room Temperature," *Appl. Phys. Lett.*, **46**, 508 (1985).

74. M. Tsuchiya, H. Sakaki, and J. Yoshino, "Room Temperature Observation of Differential Negative Resistance in an AlAs/GaAs/AlAs Resonant Tunneling Diode," *Jpn. J. Appl. Phys.*, **24**, L466 (1985).

75. S. Luryi and A. Zaslavsky, "Quantum-Effect and Hot-Electron Devices," in S. M. Sze, Ed, *Modern Semiconductor Device Physics*, Wiley, New York, 1998.

76. B. Ricco and M. Y. Azbel, "Physics of Resonant Tunneling: The One-Dimensional Double-Barrier Case," *Phys. Rev. B*, **29**, 1970 (1984).

77. S. Luryi, "Frequency Limit of Double-Barrier Resonant-Tunneling Oscillators," *Appl. Phys. Lett.*, **47**, 490 (1985).

78. E. R. Brown, J. R. Soderstrom, Jr., C. D. Parker, L. J. Mahoney, K. M. Molvar, and T. C. McGill, "Oscillations up to 712 GHz in InAs/AlSb Resonant-Tunneling Diodes," *Appl. Phys. Lett.*, **58**, 2291 (1991).

79. E. Ozbay, D. M. Bloom, and S. K. Diamond, "Looking for High Frequency Applications of Resonant Tunneling Diodes: Triggering," in L. L. Chang, E. E. Mendez, and C. Tejedor, Eds., *Resonant Tunneling in Semiconductors*, Plenum Press, New York, 1991.

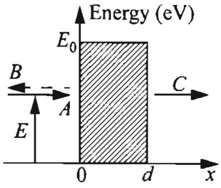
80. A. C. Seabaugh, Y. C. Kao, and H. T. Yuan, "Nine-State Resonant Tunneling Diode Memory," *IEEE Electron Dev. Lett.*, **EDL-13**, 479 (1992).

81. K. K. Ng, *Complete Guide to Semiconductor Devices*, 2nd Ed., Wiley/IEEE Press, New York, 2002.

82. F. Capasso, S. Sen, A. Y. Cho, and A. L. Hutchinson, "Resonant Tunneling Spectroscopy of Hot Minority Electrons Injected in Gallium Arsenide Quantum Wells," *Appl. Phys. Lett.*, **50**, 930 (1987).

PROBLEMS

1. Find the transmission coefficient of an electron tunneling through an one-dimensional rectangular barrier with a barrier height E_0 and a barrier width d. What is the limiting value of this coefficient if the product $\beta d \gg 1$ where $\beta \equiv \sqrt{2m^*(E_0 - E)/\hbar^2}$?
   Note: The transmission coefficient is defined as $(C/A)^2$ where A and C are the amplitudes for the incident and the transmitted wave function, respectively.



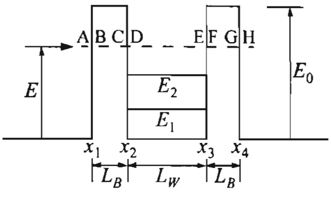
2. The I-V characteristics of a specially designed GaSb tunnel diode can be expressed by Eq. 29, with $J_P = 10^3$ A/cm^2, $V_P = 0.1$ V, $J_0 = 10^{-5}$ A/cm^2, and $J_V = 0$. The tunnel diode has a cross-sectional area of 10^{-5} cm^2. Find the largest negative differential resistance and the corresponding voltage.

3. A GaSb tunnel diode has a lead inductance of 0.1 nH, a series resistance of 4 Ω, a junction capacitance of 77 fF, and a negative resistance of −20 Ω. Find the frequency at which the real part of the input impedance becomes zero.

4. Find the speed index for the GaAs tunnel diode shown in Fig. 13. The device area is 10^{-7} cm^2, the diode is doped to 10^{20} cm^{-3} on both sides, the degeneracies on both sides are 30 mV. (*Hint*: Use abrupt junction approximation.)

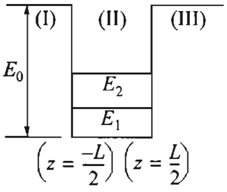
5. Molecular beam epitaxy interfaces are typically abrupt to within one or two monolayers (one monolayer ≈ 2.8 Å in GaInAs), due to terrace formation in the growth plane. Estimate the energy level broadening for the ground and first excited electron states of a 15-nm GaInAs quantum well bound by thick AlInAs barriers.
   (*Hint*: Assume the case of two monolayer thickness fluctuation and an infinite deep QW. The electron effective mass in GaInAs is $0.0427m_0$).

6. Derive the transmission coefficient for a symmetric double-barrier resonant-tunneling diode, assuming that the effective mass is constant throughout the double-barrier structure. The points A–H are adjacent to the potential steps and set boundary conditions for solution.



7. Find the lowest four resonant energy levels for a symmetric double-barrier structure with $L_B = 2$ nm, $L_W = 2$ nm, $E_0 = 3.1$ eV, and $m^* = 0.42m_0$.

8. Solve the finite-potential-well problem to find the bound energy levels $E_n < E_0$ and wavefunctions $X_n(z)$ as a function of well width L, barrier height E_0 and particle mass m^* for the symmetric quantum well. Find the number of levels contained in a potential well with $L = 10$ nm and $E_0 = 300$ meV if the particle mass $m^* = 0.067m_0$, where m_0 is the free-electron mass. These parameters correspond approximately to electrons confined in a GaAs quantum well by $Al_{0.35}Ga_{0.65}As$ heterostructure barriers.


(I)
(II)
(III)
E_0
E_2
E_1
$\left(z = \frac{-L}{2}\right)$
$\left(z = \frac{L}{2}\right)$

9. Estimate the energy broadening ΔE_1 and ΔE_2 for the lowest two levels in a model symmetric double-barrier potential resulting from tunneling out of the well (figure in Prob. 6).

The parameters are well width L = 10 nm, barrier thickness L_B = 7 nm, barrier E_0 = 300 meV, and m^* = $0.067m_0$. To find the lifetimes, consider the electron to be a semiclassical particle bouncing back and forth inside the confining double-barrier potential with a tunneling probability of escaping from the well given by Eq. 64.

10. A symmetric GaAs/AlAs RTD has a barrier width of 1.5 nm and a well width of 3.39 nm. When this RTD is inserted in the base of an HBT with an emitter flux centered on the first excited level of the RTD. Find the cutoff frequency of the HBT with the inserted RTD, if the original f_T was 100 GHz.
[*Hint*: The transit time across the RT structure is given by $(d/\upsilon_G) + (2\hbar/\Gamma)$ where d is the width of the RT structure, υ_G is the electron group velocity (10^7 cm/s), and Γ is the resonant width (20 meV).]

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

9

IMPATT Diodes

9.1 INTRODUCTION
9.2 STATIC CHARACTERISTICS
9.3 DYNAMIC CHARACTERISTICS
9.4 POWER AND EFFICIENCY
9.5 NOISE BEHAVIOR
9.6 DEVICE DESIGN AND PERFORMANCE
9.7 BARITT DIODE
9.8 TUNNETT DIODE

9.1 INTRODUCTION

The IMPATT (*impact-ionization avalanche transit-time*) diodes employ both impact-ionization and transit-time properties of semiconductor structures to produce dynamic negative resistance at microwave frequencies. Note that this negative resistance is different from, for example, that of the tunnel diode whose *I-V* curve has a negative dI/dV region. Instead, the negative resistance comes from time domain in which the ac current and voltage components are out of phase ($\tilde{V} \cdot \tilde{I}$ = negative). There are two delays that cause the current to lag behind the voltage. One is the *avalanche delay* caused by the finite buildup time of the avalanche current; the other is the *transit-time delay* from the finite time for the carriers to cross the *drift* region. When these two delays add up to half-cycle period, the diode dynamic resistance is negative at the corresponding frequency.

The negative resistance arising from transit time in semiconductor diodes was first considered by Shockley in 1954, but based on a different injection mechanism—a forward-bias *p-n* junction current.[1] In 1958, Read proposed a diode structure consisting of an avalanche region as the injection mechanism, situated at one end of a relatively high-resistance region serving as the transit-time drift space for the generated charge carriers (i.e., p^+-n-i-n^+ or n^+-p-i-p^+).[2] The experimental observation of the IMPATT oscillation was first reported by Johnston, DeLoach, and Cohen in 1965 from a regular *p-n* junction silicon diode, mounted in a microwave cavity and biased into reverse avalanche breakdown.[3,4] Oscillation based on the Read diode was reported by Lee et al. later in the same year.[5] The small-signal theory developed by Misawa[6] and by Gilden and Hines[7] has confirmed that a negative resistance of the IMPATT nature can be obtained from a *p-n* junction diode or a metal-semiconductor contact with any doping profile.

The IMPATT diode is now one of the most powerful solid-state sources of microwave frequency. At the present time, the IMPATT diode can generate the highest cw power output at millimeter-wave frequencies of all solid-state devices, from 30 GHz to above 300 GHz. But there are two noteworthy difficulties in IMPATT circuit applications: (1) The noise is high and sensitive to operating conditions; and (2) large reactances are present, which are strongly dependent on oscillation amplitude and it requires unusual care in circuit design to avoid detuning or even burnout of the diode.[8]

9.2 STATIC CHARACTERISTICS

An IMPATT diode consists of a high-field avalanche region plus a drift region. The basic members of the IMPATT diode family are shown in Fig. 1. These are the Read diode, the one-sided abrupt *p-n* junction, the *p-i-n* diode (Misawa diode), the two-sided (double-drift) diode, the hi-lo and lo-hi-lo diodes (modified Read diodes).

We shall now consider their static characteristics, such as the field distribution, breakdown voltage, and space-charge effect. Consider Fig. 1a first which shows the doping profile, electric-field distribution, and ionization integrand at breakdown condition for an idealized Read diode (p^+-n-i-n^+ or its dual n^+-p-i-p^+). The middle *n*- and *i*-regions are totally depleted, indicated by the shaded area. The ionization integrand is given by

$$\langle\alpha\rangle \equiv \alpha_n \exp\left[-\int_x^{W_D}(\alpha_n-\alpha_p)dx'\right] \qquad \alpha_n > \alpha_p \tag{1}$$

where α_n and α_p are the ionization rates of electrons and holes, respectively, and W_D is the depletion width.

The avalanche breakdown condition as discussed in Chapter 2 is given by

$$\int_0^{W_D}\langle\alpha\rangle dx = 1 . \tag{2}$$

Because of the strong dependence of α on an electric field, we note that the *avalanche region* is highly localized, that is, most of the multiplication processes occur in a narrow region near the highest field between 0 and x_A, where x_A is defined as the width of the avalanche region (to be discussed later). The voltage drop across the ava-

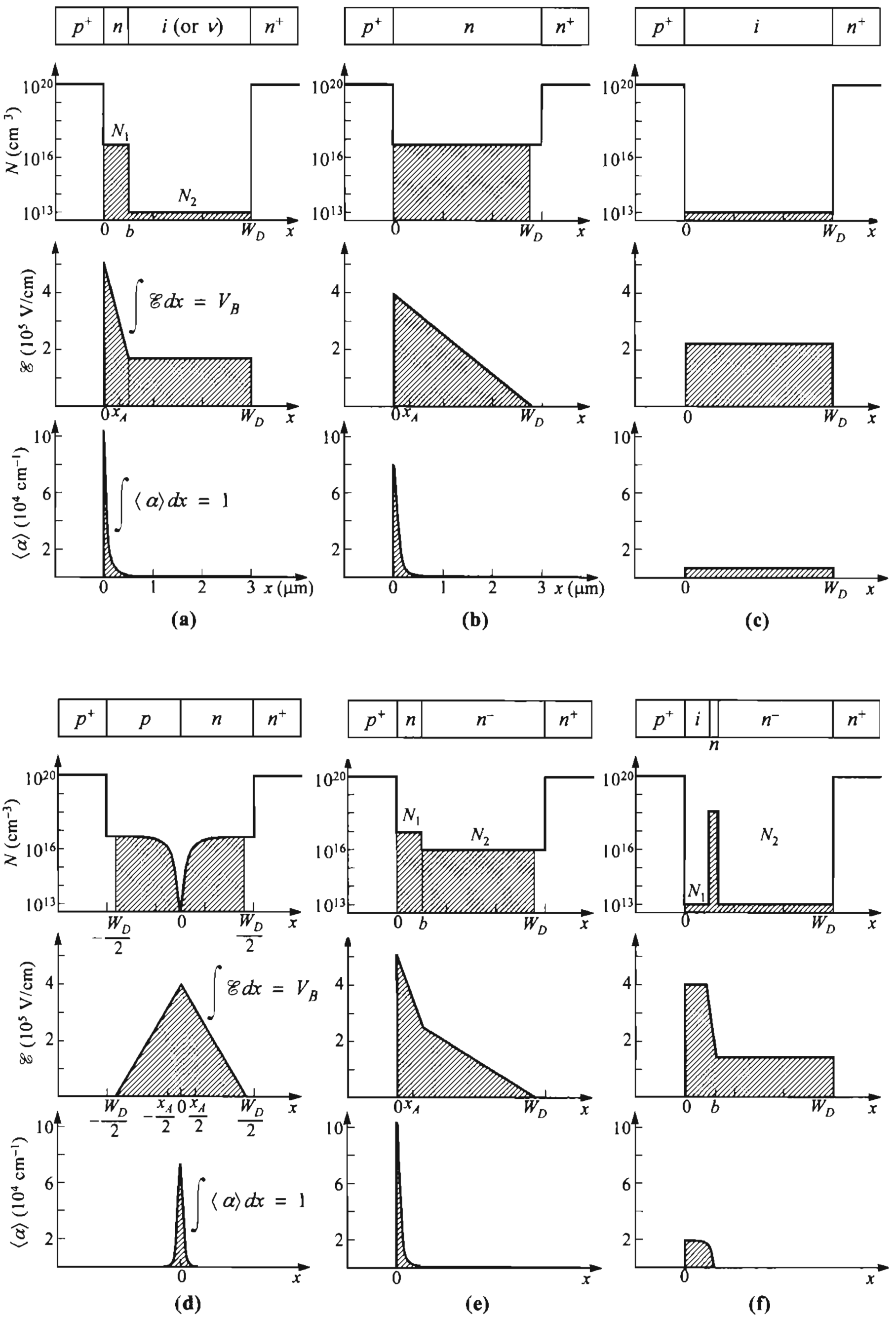


Fig. 1 Doping profile, electric-field distribution, and ionization integrand for (a) Read diode; (b) one-sided abrupt diode; (c) *p-i-n* diode; (d) double-drift diode; (e) hi-lo structure; and (f) lo-hi-lo structure.

lanche region x_A is defined as V_A. It will be shown that both x_A and V_A have a profound effect on the optimum current density and the maximum efficiency of an IMPATT diode. The layer outside the avalanche region ($x_A \leq x \leq W_D$) is called the drift region.

There are two limiting cases of the Read doping profiles. As the N_2-region becomes zero, we have a one-sided abrupt p^+n junction. Figure 1b describes the structure of a one-sided abrupt *p-n* junction. The avalanche region is highly localized near the junction. On the other hand, when the N_1-region becomes zero instead we have a *p-i-n* diode (Fig. 1c).[6] The *p-i-n* diode has a uniform field across the intrinsic layer under low-current conditions. The avalanche region corresponds to the full intrinsic layer width. Figure 1d describes the structure of a two-sided abrupt *p-n* junction. The avalanche region is located near the center of the depletion layer. The slight asymmetry of the integrand $\langle \alpha \rangle$ with respect to the location of the maximum field is because of the large difference between α_n and α_p in Si. If $\alpha_n \approx \alpha_p$ as in the case of GaP, $\langle \alpha \rangle$ reduces to

$$\langle \alpha \rangle = \alpha_n = \alpha_p \tag{3}$$

and the avalanche region is symmetrical with respect to $x = 0$.

Figure 1e shows a modified Read diode, the hi-lo structure, in which the doping N_2 is larger than that for a Read diode.[9] Figure 1f shows another modified Read diode, the lo-hi-lo structure, in which a clump of charge is located at $x = b$. Since a nearly uniform high-field region exists from $x = 0$ to $x = b$, the value of the maximum field can be much lower than that for a hi-lo diode.

9.2.1 Breakdown Voltage

The breakdown voltage for the one-sided abrupt junction has been considered in Chapter 2. We can use the same method as outlined in that chapter to calculate the breakdown voltages of other diodes. Even though the breakdown is ultimately determined by the ionization integrand, it is helpful and simpler to predict breakdown based on the maximum field that has been calculated at breakdown condition. Notice that in some of the structures in Fig. 1 (a, c, f), the maximum depletion is terminated by the lightly doped width and there is a discontinuity of field near the n^+-terminal. In all others, the depletion-width edge is determined mostly by the doping and the field drops to zero at the depletion edge.

For the one-sided (Fig. 1b) and two-sided symmetrical abrupt junctions (Fig. 1d), the breakdown voltages are given respectively by

$$V_B = \frac{1}{2}\mathscr{E}_m W_D = \frac{\varepsilon_s \mathscr{E}_m^2}{2qN} \qquad \text{(1-sided)}, \tag{4a}$$

$$V_B = \frac{1}{2}\mathscr{E}_m W_D = \frac{\varepsilon_s \mathscr{E}_m^2}{qN} \qquad \text{(2-sided)}, \tag{4b}$$

where $\mathscr{E}_m$ is the maximum field, which occurs at $x = 0$. The maximum fields at breakdown for Si and $\langle 100 \rangle$-oriented GaAs two-sided (symmetrical) abrupt junctions, together with the one-sided abrupt junctions, are shown in Fig. 2. Once the doping is known, the breakdown voltage can be calculated from Eq. 4a or 4b, using the

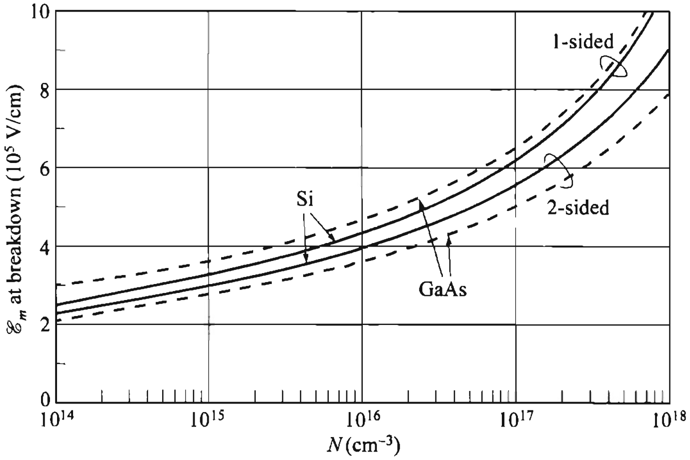


Fig. 2 Maximum electric field at breakdown vs. doping for Si and GaAs one-sided and two-sided abrupt junctions. (After Refs. 10 and 11.)

maximum field value from Fig. 2. The applied reverse voltage at breakdown is equal to $(V_B - \psi_{bi})$, where ψ_{bi} is the built-in potential, in the case of symmetrical abrupt junction given by $2(kT/q)\ln(N/n_i)$. Usually ψ_{bi} is negligible for practical IMPATT diodes.

For the Read diode, the breakdown voltage is given by

$$V_B = \mathscr{E}_m W_D - \frac{qN_1 b}{\varepsilon_s}\left(W_D - \frac{b}{2}\right). \tag{5}$$

The depletion width is limited by the thickness of the n^--layer. For the hi-lo diode, the breakdown and depletion width are given by

$$V_B = \frac{\mathscr{E}_m}{2}(W_D + b) - \frac{qN_1 W_D b}{2\varepsilon_s} = \frac{\mathscr{E}_m b}{2} + \frac{qN_2 W_D (W_D - b)}{2\varepsilon_s}, \tag{6}$$

$$W_D = \frac{\varepsilon_s \mathscr{E}_m}{qN_2} - b\left(\frac{N_1}{N_2} - 1\right). \tag{7}$$

The maximum field at breakdown for a Read diode or a hi-lo diode with a given N_1 is found to be essentially the same (within 1%) as the value of the one-sided abrupt junction with the same N_1, provided that the avalanche width x_A is smaller than b.[12] Therefore, the breakdown voltages can be calculated from Eqs. 5 and 6 using the maximum field value of Fig. 2.

For the lo-hi-lo diode with a narrow fully depleted clump of charge, the breakdown voltage is given by

$$V_B = \mathscr{E}_m b + \left(\mathscr{E}_m - \frac{qQ}{\varepsilon_s}\right)(W_D - b), \tag{8}$$

where Q is the impurity density per area (number/cm^2) in the clump. Since the maximum field is nearly constant for $0 \leq x \leq b$, $\langle \alpha \rangle$ is equal to $1/b$ at breakdown. The maximum field $\mathscr{E}_m$ can be calculated from the field-dependent ionization coefficient.

9.2.2 Avalanche Region and Drift Region

The avalanche region of an ideal *p-i-n* diode is the full intrinsic-layer width. For the Read diode and *p-n* junctions, however, the region of carrier multiplication is restricted to a narrow region close to the metallurgical junction. The contribution to the integral in Eq. 2 decreases rapidly as x departs from the metallurgical junction. Thus a reasonable definition of the avalanche-region width x_A is obtained by taking the distance over which 95% of the contribution to the integral is obtained, that is,

$$\int_0^{x_A} \langle \alpha \rangle dx \quad \text{or} \quad \int_{-x_A/2}^{x_A/2} \langle \alpha \rangle dx \quad = 0.95\,. \tag{9}$$

Figure 3 shows the avalanche widths as a function of the doping for Si and GaAs diodes.[11] Also shown are the depletion widths of Si and GaAs symmetrical two-sided junctions. For a given doping, the Si n^+-p junction has a narrower avalanche width than that in p^+-n junction because of the difference in ionization rates ($\alpha_n > \alpha_p$). For a Read diode or a hi-lo diode, the avalanche region will be the same as a one-sided

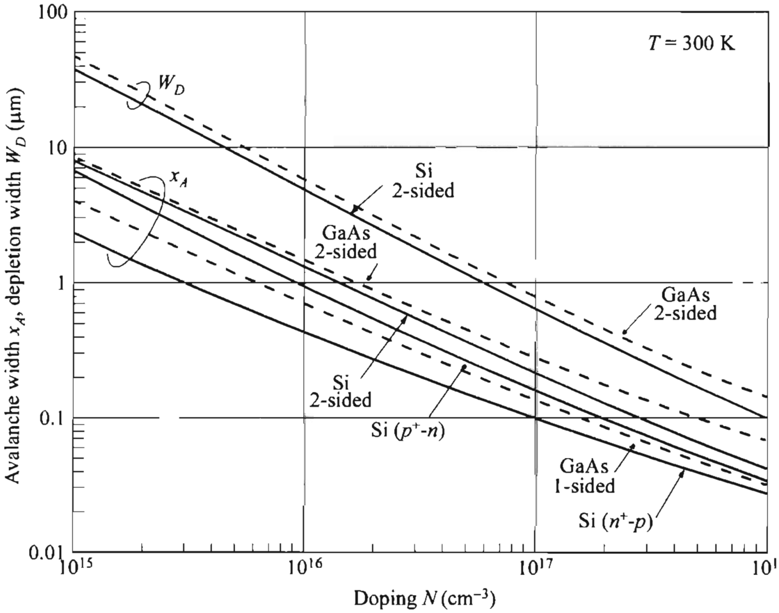


Fig. 3 Avalanche region widths x_A for Si and GaAs junctions. Also shown are depletion widths W_D of Si and GaAs symmetrical 2-sided junctions. (After Ref. 11.)

abrupt junction with the same doping N_1. For a lo-hi-lo diode, the avalanche region width is equal to the distance between the metallurgical junction and the charge clump $x_A = b$.

The drift region is the depletion layer excluding the avalanche region, or $x_A \leq x \leq W_D$. The most-important parameter in the drift region is the carrier drift velocity. To obtain consistent and predictable carrier transit time across the drift region, the electric field in this region should be high enough that the generated carriers can travel at their saturation velocities υ_s. For silicon the electric field should be larger than 10^4 V/cm. For GaAs, the field can be much smaller ($\approx 10^3$ V/cm) due to its high carrier mobilities.

For *p-i-n* diodes, this requirement is fulfilled automatically, because at breakdown the field (which is approximately constant over the full intrinsic width) is much larger than the required field for velocity saturation. For a Read diode the minimum field in the drift region is given by

$$\mathscr{E}_{\min} = \mathscr{E}_m - \frac{q[N_1 b + N_2(W_D - b)]}{\varepsilon_s}. \tag{10}$$

Clearly, from the previous discussion, a Read diode can be designed so that $\mathscr{E}_{\min}$ is sufficiently large. For abrupt junctions, since the field drops to zero at the depletion edge, some regions always have fields smaller than the minimum required field. The low-field region, however, constitutes only a small percent of the total depletion region. For example, for a Si p^+-n junction with 10^{16} cm^{-3} background doping, the maximum field at breakdown is 4×10^5 V/cm. The ratio of the low-field region (for a field less than 10^4 V/cm) to the total depletion layer is $10^4/4 \times 10^5 = 2.5\%$. For a GaAs p^+-n junction with the same doping, the low-field region is less than 0.2%. Thus, the low-field region has negligible effect on the total carrier transit time across the depletion layer.

9.2.3 Temperature and Space-Charge Effects

The breakdown voltages and the maximum electric fields discussed previously are calculated for room temperature under isothermal conditions, free of space-charge effects (from high-level injection), and in the absence of oscillation. Under operating conditions, however, the IMPATT diode is biased well into avalanche breakdown, and the current density is usually very high. This results in a considerable temperature rise in the junction and a large space-charge effect.

The ionization rates of electrons and holes decrease with increasing temperature.[13] Thus for an IMPATT diode with a given doping profile, the breakdown voltage will increase with increasing temperature. As the dc power (product of reverse voltage and reverse current) increases, both the junction temperature and the breakdown voltage increase. Eventually, the diode fails to operate, mainly because of permanent damage that results from excessive heating in localized spots. Thus, the rising temperature of the junction imposes a severe limit on device operation. To prevent the temperature rise, one must use a suitable heat sink. This will be considered in Section 9.4.4.

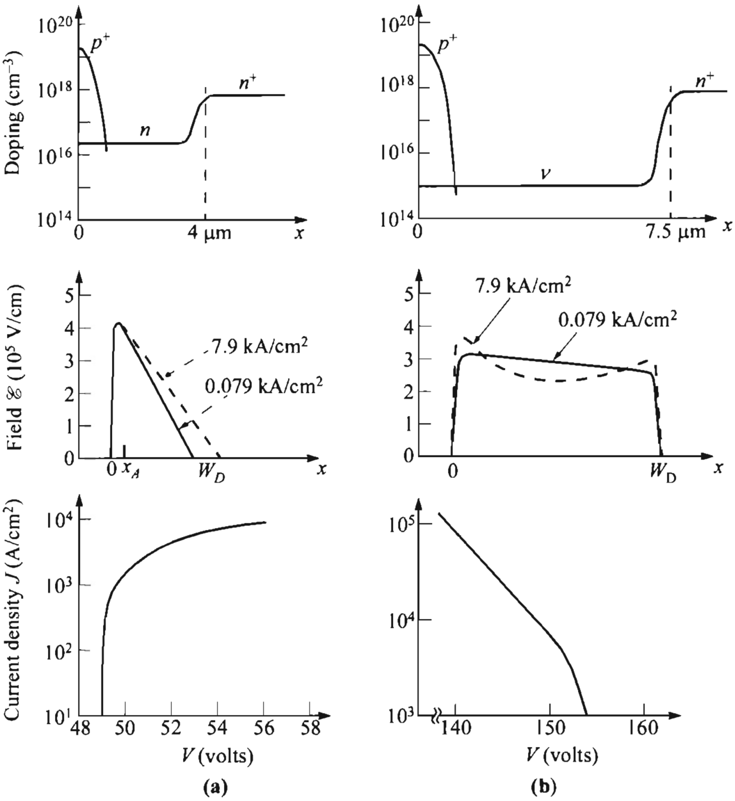
The space-charge effect causes a change of electric field in the depletion region due to extra space charge. This effect gives rise to a positive dc differential resistance for abrupt junctions and a negative dc differential resistance for *p-i-n* diodes.[14]

Consider first a one-sided p^+-n-n^+ abrupt junction as shown in Fig. 4a. When the applied voltage is equal to the breakdown voltage V_B, the electric field $\mathscr{E}(x)$ has its maximum absolute value $\mathscr{E}_m$ at $x = 0$. If we assume that the electrons travel at their saturation velocity v_s across the depletion region, the space-charge-limited current is given by

$$I = Aq\Delta n v_s \tag{11}$$

where Δn is the high-level injected carrier density and A the area. The disturbance $\Delta\mathscr{E}(x)$ in the electric field due to the space charge is obtained from Eq. 11 and the Poisson equation:

$$\Delta\mathscr{E}(x) \approx \frac{Ix}{A\varepsilon_s v_s}. \tag{12}$$



(a) **(b)**

Fig. 4 Doping profile, field, and current-voltage characteristics of (a) p^+-n-n^+ and (b) p^+-ν-n^+ diodes. Area = 10^{-4} cm^2. (After Ref. 14.)

If we assume that all the carriers are generated within the avalanche width x_A, the disturbance in voltage caused by the carriers in the drift region $(W_D - x_A)$ is obtained by integrating $\Delta\mathscr{E}(x)$ over this drift region:

$$\Delta V_B \approx \int_0^{W_D - x_A} \frac{Ix}{A\varepsilon_s v_s} dx \approx I\frac{(W_D - x_A)^2}{2A\varepsilon_s v_s}. \tag{13}$$

The total applied voltage is, thus, increased by this amount to maintain the same current. The space-charge resistance[15] is obtained from Eq. 13:

$$R_{SC} \equiv \frac{\Delta V_B}{I} \approx \frac{(W_D - x_A)^2}{2A\varepsilon_s v_s}. \tag{14}$$

For our example shown in Fig. 4a, the space-charge resistance is about 20 Ω.

For a *p-i-n* or a *p-ν-n* diode, the situation is different from that of a p^+-*n* junction. When the applied reverse voltage is just large enough to cause avalanche breakdown, the reverse current is small. The space-charge effect can be neglected and the electric field is essentially uniform across the depletion layer. As the current increases, more electrons are generated near the p^+-ν boundary and more holes are generated near the ν-n^+ boundary (by impact ionization as the electric field has double peaks, Fig. 4b). These charges will cause a reduction of the field in the center of the ν-region that decreases the total terminal voltage. This reduction results in a negative differential dc resistance for the *p-ν-n* diode, as shown in Fig. 4b.

9.3 DYNAMIC CHARACTERISTICS

9.3.1 Injection Phase Delay and Transit-Time Effect

We consider first the injection phase delay and transit-time effect of an idealized device,[16] whose structure is shown in Fig. 5 where we move the x-origin to the right of the avalanche region (plane of charge injection). The terminal voltage and the avalanche generation rate are also shown in relation to each other. The terminal voltage, of angular frequency ω, has a mean value at the verge of avalanche breakdown V_B. In the positive cycle, avalanche multiplication begins. The generation rate of carriers, however, as shown, is not in unison with the voltage or field. This is because the generation rate is not only a function of the field but also of the number of existing carriers. After the field passes the peak value, the generation rate continues to grow until the field is below the critical value. This phase lag is approximately π and is called the injection phase delay.

Assume that an avalanche charge pulse is injected at $x = 0$, in Fig. 5, with a given phase angle delay ϕ with respect to the terminal voltage. Also assume that the applied dc voltage across the diode causes the carriers to travel at the saturation velocity v_s in the drift region, $0 \le x \le W_D$. The ac conduction current density $\tilde{J}_c$ is also a function of location x, and its magnitude is related to the total ac current density by:

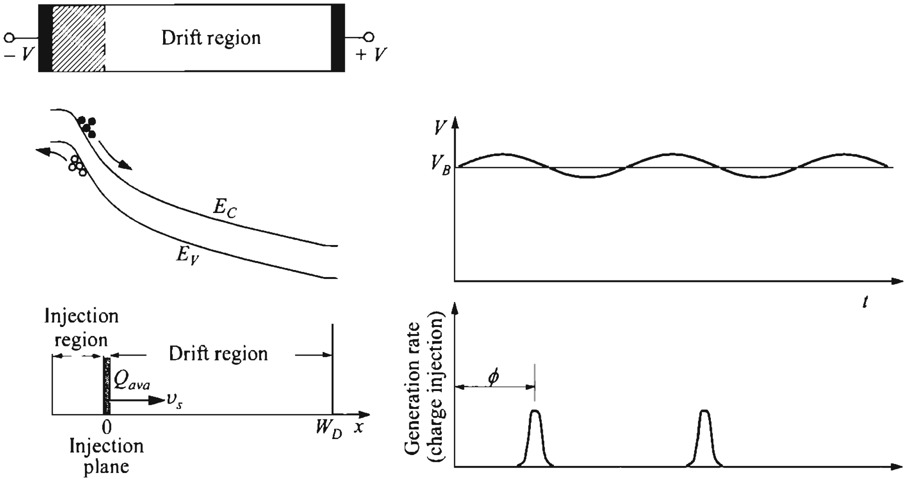

Drift region
−V
+V
E_C
E_V
Injection region
Drift region
Q_{ava}
v_s
0
Injection plane
W_D
x
(a)
V
V_B
t
Generation rate (charge injection)
ϕ
t
(b)

Fig. 5 (a) Idealized IMPATT diode with carrier injection at $x = 0$ and a drift region with saturation velocity. (b) Terminal voltage and avalanche generation rate in time domain. The avalanche lags the voltage by $\phi \approx \pi$.

$$\tilde{J}_c(x) = \tilde{J}\exp\left[-j\left(\phi + \frac{\omega x}{v_s}\right)\right]. \tag{15}$$

The total ac current anywhere in the drift region is given by the sum of the conduction current and the displacement current:

$$\begin{aligned}\tilde{J}(x) &= \tilde{J}_c(x) + \tilde{J}_d(x)\\ &= \tilde{J}\exp\left[-j\left(\phi + \frac{\omega x}{v_s}\right)\right] + j\omega\varepsilon_s\tilde{\mathscr{E}}(x)\end{aligned} \tag{16}$$

where $\tilde{\mathscr{E}}(x)$ is the ac field. From Eqs. 15 and 16 we obtain

$$\tilde{\mathscr{E}}(x) = \frac{\tilde{J}(x)}{j\omega\varepsilon_s}\left\{1 - \exp\left[-j\left(\phi + \frac{\omega x}{v_s}\right)\right]\right\}. \tag{17}$$

Integrating Eq. 17 gives the ac impedance

$$Z \equiv \frac{1}{\tilde{J}}\int_0^{W_D} \tilde{\mathscr{E}}(x)dx = \frac{1}{j\omega C_D}\left\{1 - \frac{\exp(-j\phi)[1 - \exp(-j\theta)]}{j\theta}\right\}, \tag{18}$$

where C_D is the depletion capacitance per unit area ε_s/W_D, and θ is the transit angle

$$\theta = \frac{\omega W_D}{v_s}. \tag{19}$$

By taking the real and the imaginary parts of Eq. 18, we obtain

$$R_{ac} = \frac{\cos\phi - \cos(\phi + \theta)}{\omega C_D \theta}, \tag{20}$$

$$X = -\frac{1}{\omega C_D} + \frac{\sin(\phi + \theta) - \sin\phi}{\omega C_D \theta}. \tag{21}$$

We consider next the influence of the injection phase ϕ on the ac resistance R_{ac}, based on Eq. 20. When ϕ equals zero (no phase delay), the resistance is proportional to $(1 - \cos\theta)/\theta$, which is always greater or equal to zero, as shown in Fig. 6a; that is, there is no negative resistance. Therefore, the transit-time effect alone cannot give rise to negative resistance. However, for any nonzero ϕ, the resistance is negative for certain transit angles. For example, at $\phi = \pi/2$, the largest negative resistance occurs near $\theta = 3\pi/2$, as shown in Fig. 6b. For $\phi = \pi$, the same occurs near $\theta = \pi$, as shown in Fig. 6c. This corresponds to the IMPATT operation, in which the buildup of the injection current due to impact avalanche introduces a phase delay of about π, and the transit time in the drift region gives an additional π delay.

The foregoing considerations have confirmed the importance of the injection delay. The problem of finding active transit-time devices has been reduced to finding a means to delay the injection of conduction current into the drift region. From Fig. 6 we observe that the sum of the injection phase and the optimum transit angle, $\phi + \theta_{opt}$,

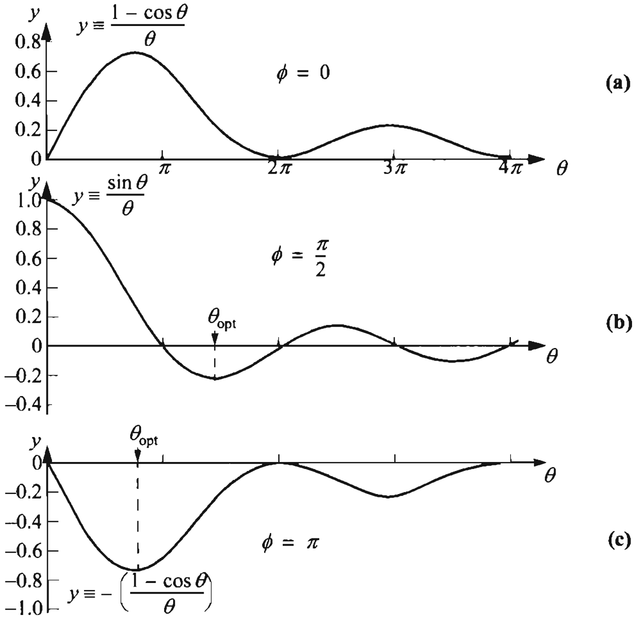


Fig. 6 AC resistance vs. transit angle for three different injection phase delays. (a) $\phi = 0$. (b) $\phi = \pi/2$. (c) $\phi = \pi$.

is approximately equal to 2π. As ϕ increases from zero, the magnitude of the negative resistance becomes larger.

9.3.2 Small-Signal Analysis

The small-signal analysis was first considered by Read[2] and developed further by Gilden and Hines.[7] For simplicity we assume that $\alpha_n = \alpha_p = \alpha$, and that the saturation velocities of holes and electrons are equal. Figure 7a shows the model of a Read diode. According to the discussion in Section 9.2, we have divided the diode into three regions: (1) the avalanche region, which is assumed to be thin so that space-charge and signal delay can be neglected; (2) the drift region, where no carriers are

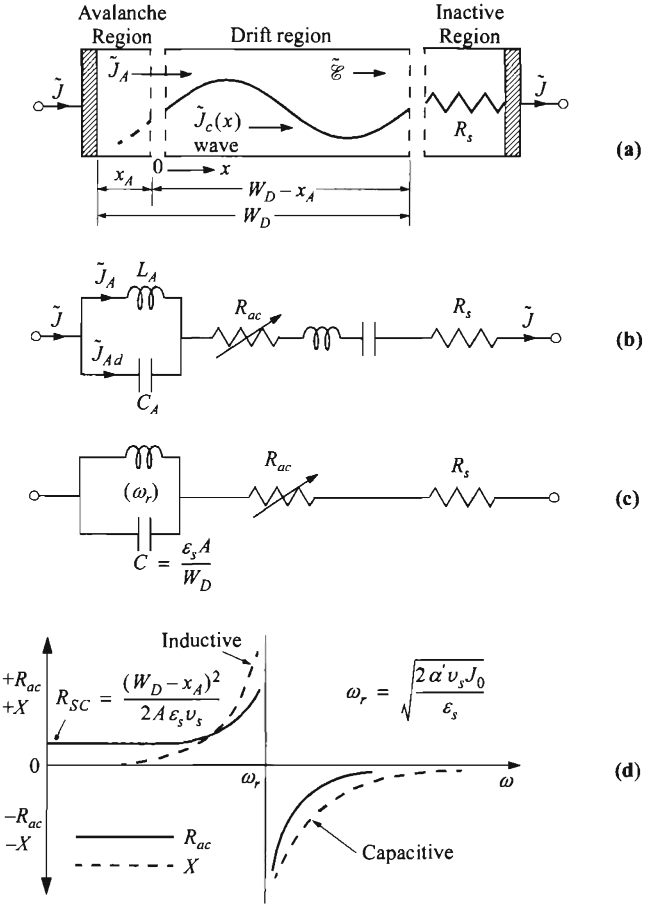

Avalanche Region
Drift region
Inactive Region
$\tilde{J}_A$
$\tilde{\mathscr{E}}$
$\tilde{J}$
$\tilde{J}_c(x)$
wave
R_s
$\tilde{J}$
x_A
0
x
$W_D - x_A$
W_D
(a)
$\tilde{J}_A$
L_A
$\tilde{J}$
$\tilde{J}_{Ad}$
C_A
R_{ac}
R_s
$\tilde{J}$
(b)
(ω_r)
R_{ac}
R_s
$C = \frac{\varepsilon_s A}{W_D}$
(c)
Inductive
$+R_{ac}$
$+X$
$R_{SC} = \frac{(W_D - x_A)^2}{2A\varepsilon_s v_s}$
$\omega_r = \sqrt{\frac{2\alpha' v_s J_0}{\varepsilon_s}}$
0
ω_r
ω
$-R_{ac}$
$-X$
R_{ac}
X
Capacitive
(d)

Fig. 7 (a) Model of Read diode with avalanche, drift, and inactive regions, and (b) its equivalent circuit. For small transit angle θ_d, (c) equivalent circuit and (d) plots of real and imaginary parts of impedance vs. angular frequency ω. (After Ref. 7.)

generated, and all carriers entering from the avalanche region travel at their saturation velocities; and (3) an inactive region that adds undesirable parasitic resistance.

The two active regions interact with each other, because the ac electric field is continuous across the boundary between them. We shall use a "0" subscript to indicate dc quantities, and "~" to indicate small-signal ac quantities. For quantities including both dc and ac components, no "0" subscript or "~" will be added. We first define $\tilde{J}_A$ as the avalanche current density, which is the ac conduction (particle) current in the avalanche region, and $\tilde{J}$ as the total ac current density. With our assumption of a thin avalanche region, $\tilde{J}_A$ is presumed to enter the drift region without delay. With the assumption of a saturation velocity υ_s, the ac conduction current $\tilde{J}_c(x)$ in the drift region propagates as an unattenuated wave (with only phase change) at this drift velocity,

$$\begin{aligned}\tilde{J}_c(x) &= \tilde{J}_A \exp\left(\frac{-j\omega x}{\upsilon_s}\right)\\ &\equiv \gamma \tilde{J} \exp\left(\frac{-j\omega x}{\upsilon_s}\right),\end{aligned} \tag{22}$$

where $\gamma = \tilde{J}_A/\tilde{J}$ is the complex fraction relating the avalanche current to the total current. At any x-location, the total alternating current $\tilde{J}$ equals the sum of the conduction current $\tilde{J}_c$ and the displacement current $\tilde{J}_d$. This sum is constant, independent of position x. Equation 17 is rewritten as (putting $\phi = 0$):

$$\tilde{\mathscr{E}}(x) = \frac{\tilde{J}}{j\omega\varepsilon_s}\left[1 - \gamma \exp\left(-\frac{j\omega x}{\upsilon_s}\right)\right]. \tag{23}$$

Integrating $\tilde{\mathscr{E}}(x)$ gives the voltage across the drift region in terms of $\tilde{J}$. The coefficient γ is derived in the analysis that follows.

Avalanche Region. Consider first the avalanche region. Under the dc condition, the direct current J_0 $(= J_{po} + J_{no})$ is related to the thermally generated reverse saturation current J_s $(= J_{ns} + J_{ps})$ by

$$J_0 = \frac{J_s}{1 - \int_0^{W_D} \langle\alpha\rangle dx}. \tag{24}$$

At breakdown, J_0 approaches infinity and the integral is equal to unity. In the dc case the integral cannot be greater than unity. This is not necessarily so for a rapidly varying field. The differential equation for the current as a function of time will now be derived. Under the conditions that (1) electrons and holes have equal ionization rates and equal saturation velocities, and (2) the drift current components are much larger than the diffusion component, the basic device equations in the one-dimensional case can be written as follows:

$$J = J_n + J_p = q\upsilon_s(n + p), \qquad \text{current-density equation}, \tag{25}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} + \alpha v_s(n+p)\ , \qquad \text{continuity equations}, \tag{26a}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} + \alpha v_s(n+p)\ . \tag{26b}$$

The second terms on the right-hand side of Eqs. 26a and 26b correspond to the generation rate of the electron-hole pairs by avalanche multiplication. This generation rate is so large compared to the rate of thermal generation that the latter can be neglected. Adding Eqs. 26a and 26b and integrating from $x = 0$ to x_A gives

$$\tau_A\frac{dJ}{dt} = -(J_p - J_n)\Big|_0^{x_A} + 2J\int_0^{x_A} \alpha dx \tag{27}$$

where $\tau_A = x_A/v_s$ is the transit time across the multiplication region. The boundary conditions are that the electron current at $x = 0$ consists entirely of the reverse saturation current J_{ns}. Thus at $x = 0$, the boundary condition is

$$J_p - J_n = -2J_n + J = -2J_{ns} + J. \tag{28a}$$

At $x = x_A$ the hole current consists of the reverse saturation current J_{ps} generated in the space-charge region, so we have

$$J_p - J_n = 2J_p - J = 2J_{ps} - J. \tag{28b}$$

With these boundary conditions, Eq. 27 becomes

$$\frac{dJ}{dt} = \frac{2J}{\tau_A}\left(\int_0^{x_A} \alpha dx - 1\right) + \frac{2J_s}{\tau_A}. \tag{29}$$

In the dc case, J is the direct current J_0, so that Eq. 29 reduces to Eq. 24.

We now simplify Eq. 29 by substituting $\bar{\alpha}$ in place of α, where $\bar{\alpha}$ is an average value of α obtained by evaluating the integral over the extent of the avalanche region. We obtain (by neglecting the term J_s)

$$\frac{dJ}{dt} = \frac{2J}{\tau_A}(\bar{\alpha}x_A - 1). \tag{30}$$

Furthermore the small-signal assumptions are now made:

$$\bar{\alpha} = \bar{\alpha}_0 + \tilde{\alpha}\exp(j\omega t) \approx \bar{\alpha}_0 + \bar{\alpha}'\tilde{\mathscr{E}}_A\exp(j\omega t), \tag{31}$$

$$\bar{\alpha}x_A = 1 + x_A\bar{\alpha}'\tilde{\mathscr{E}}_A\exp(j\omega t), \tag{32}$$

$$J = J_0 + \tilde{J}_A\exp(j\omega t), \tag{33}$$

$$\mathscr{E} = \mathscr{E}_0 + \tilde{\mathscr{E}}_A\exp(j\omega t), \tag{34}$$

where $\alpha' \equiv \partial\alpha/\partial\mathscr{E}$ and the substitution $\tilde{\alpha} = \alpha'\tilde{\mathscr{E}}_A$ has been employed. Substituting the expressions above into Eq. 30, neglecting products of higher-order terms, leads to the expression for the ac component of the avalanche conduction current,

$$\tilde{J}_A = \frac{2\alpha' x_A J_0 \tilde{\mathscr{E}}_A}{j\omega\tau_A}. \tag{35}$$

The displacement current in the avalanche region is simply given by

$$\tilde{J}_{Ad} = j\omega\varepsilon_s \tilde{\mathscr{E}}_A. \tag{36}$$

The above are the two components of the total circuit current in the avalanche region. For a given field, the avalanche current $\tilde{J}_A$ is reactive and varies inversely with ω as in an inductor. The other component, J_{Ad}, is also reactive and varies directly with ω as in a capacitor. Thus the avalanche region behaves as an LC parallel circuit. The equivalent circuit is shown in Fig. 7b, where the inductance and capacitance are given as (where A is the diode area)

$$L_A = \frac{\tau_A}{2J_0\alpha' A}, \tag{37}$$

$$C_A = \frac{\varepsilon_s A}{x_A}. \tag{38}$$

The resonant frequency of this combination is given by

$$\omega_r = 2\pi f_r = \sqrt{\frac{2\alpha' v_s J_0}{\varepsilon_s}}. \tag{39}$$

A thin avalanche region, therefore, behaves as an antiresonant circuit with a resonant frequency proportional to the square root of the direct current density J_0. The impedance of the avalanche region has the simple form

$$Z_A = \frac{x_A}{j\omega\varepsilon_s A}\left[\frac{1}{1-(\omega_r^2/\omega^2)}\right] = \frac{1}{j\omega C_A}\left[\frac{1}{1-(\omega_r^2/\omega^2)}\right]. \tag{40}$$

The factor γ can be expressed as

$$\gamma \equiv \frac{\tilde{J}_A}{\tilde{J}} = \frac{1}{1-(\omega^2/\omega_r^2)}. \tag{41}$$

Drift Region. Combining Eqs. 41 and 23 and integrating over the drift length $(W_D - x_A)$ gives an expression for the ac voltage across this region as,

$$\tilde{V}_d = \frac{(W_D - x_A)\tilde{J}}{j\omega\varepsilon_s}\left\{1 - \frac{1}{1-(\omega^2/\omega_r^2)}\left[\frac{1-\exp(-j\theta_d)}{j\theta_d}\right]\right\}, \tag{42}$$

where θ_d is the transit angle of the drift space

$$\theta_d \equiv \frac{\omega(W_D - x_A)}{v_s} \equiv \omega\tau_d \tag{43}$$

and

$$\tau_d = \frac{(W_D - x_A)}{v_s}. \tag{44}$$

We may also define $C_D \equiv A\varepsilon_s/(W_D - x_A)$ as the capacitance of the drift region. From Eq. 42 we obtain the impedance for the drift region,

$$\begin{aligned} Z_d \equiv \frac{\tilde{V}_d}{A\tilde{J}} &= \frac{1}{\omega C_D}\left[\frac{1}{1-(\omega^2/\omega_r^2)}\left(\frac{1-\cos\theta_d}{\theta_d}\right)\right] + \frac{j}{\omega C_D}\left[\frac{1}{1-(\omega^2/\omega_r^2)}\left(\frac{\sin\theta_d}{\theta_d}\right) - 1\right] \\ &= R_{ac} + jX \end{aligned} \tag{45}$$

where R_{ac} and X are the resistance and reactance, respectively. At low frequencies and with $\phi = 0$, it can be seen that Eq. 45 reduces to Eqs. 20 and 21. The real part (resistance) will be negative for all frequencies above ω_r, except for nulls at $\theta_d = 2\pi\times$integer. The resistance is positive for frequencies below ω_r, and approaches a finite value at zero frequency:

$$R_{ac}(\omega \rightarrow 0) = \frac{\tau_d}{2C_D} = \frac{(W_D - x_A)^2}{2A\varepsilon_s v_s}. \tag{46}$$

The low-frequency small-signal resistance is a consequence of the space charge in the finite thickness of the drift region, and the expression above is identical to Eq. 14 derived previously.

Total Impedance. The total impedance is the sum of the impedances of the avalanche region, drift region, and passive resistance R_s of the inactive region:

$$\begin{aligned} Z = {} & \frac{(W_D - x_A)^2}{2A\varepsilon_s v_s}\left[\frac{1}{1-(\omega^2/\omega_r^2)}\right]\left(\frac{1-\cos\theta_d}{\theta_d^2/2}\right) \\ & + \frac{j}{\omega C_D}\left\{\left(\frac{\sin\theta_d}{\theta_d} - 1\right) - \frac{(\sin\theta_d/\theta_d) + [x_A/(W_D - x_A)]}{1-(\omega_r^2/\omega^2)}\right\} + R_S. \end{aligned} \tag{47}$$

The real part is the dynamic resistance and it changes sign from positive to negative as ω becomes larger than ω_r.

Equation 47 has been cast in a form that can be simplified directly for the case of small transit angle θ_d. For $\theta_d < \pi/4$, Eq. 47 reduces to

$$Z = \frac{(W_D - x_A)^2}{2Av_s\varepsilon_s[1-(\omega^2/\omega_r^2)]} + \frac{j}{\omega C_D}\left[\frac{1}{(\omega_r^2/\omega^2) - 1}\right] + R_S \tag{48}$$

where $C_D \equiv \varepsilon_s A/W_D$ corresponding to the total depletion capacitance. The equivalent circuit and frequency dependence of the real and imaginary parts of the impedance are shown in Fig. 7c and d, respectively. In Eq. 48, note again that the first term is the active resistance, which becomes negative for $\omega > \omega_r$. The second term is reactive and corresponds to a parallel resonant circuit that includes the diode capacitance and a shunt inductor. The reactance is inductive for $\omega < \omega_r$ and capacitive for $\omega > \omega_r$. In other words, the resistance becomes negative at the frequency where the reactive component changes sign.

9.4 POWER AND EFFICIENCY

9.4.1 Large-Signal Operation

Under large-signal operation, a high-field avalanche region exists at the p^+-n junction of a Read diode (Fig. 1a), where electron-hole pairs are generated; and a constant-field drift region exists in the low-doped ν-region. The generated holes quickly enter the p^+-region and the generated electrons are injected into the drift region, where they do work that produces external power. As discussed before, the ac variation of the injected charge lags the ac voltage by about π, as *injection delay* ϕ shown in Fig. 8. The injected carriers then enter the drift region, where they traverse at saturation velocity, introducing the *transit-time delay*. The induced external current is also shown. Comparing the ac voltage and the external current clearly shows that the diode exhibits a negative resistance at its terminals.

For large-signal operation, the terminal current is mostly a result of the charge generated by avalanche multiplication and its movement. When the electron packet (of charge density Q_{ava}) traverses toward the n^+-region (anode) with a saturation velocity, an external current is induced. The terminal conduction current can be obtained by calculating the induced charge partitioned at the anode or cathode. Consider for example, the charge density at the anode Q_A which is a function of the location of Q_{ava}, and it is given by

$$Q_A(t) = \frac{Q_{ava}x}{W_D} = \frac{Q_{ava}v_s t}{W_D}. \tag{49}$$

The peak conduction current is thus given by

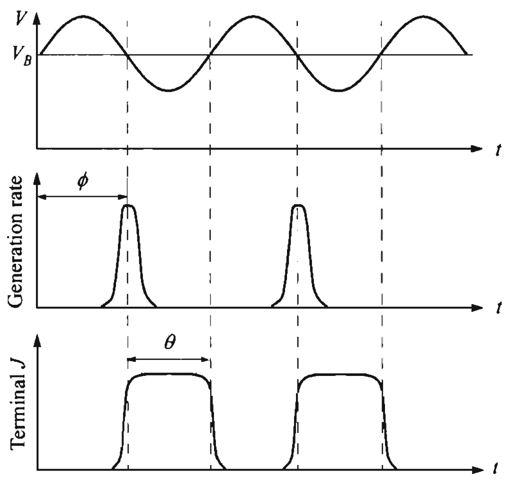


Fig. 8 Large-signal operation of IMPATT diode, showing terminal voltage, avalanche generation rate, and terminal current. ϕ= injection delay. θ= transit-time delay.

$$J_c = \frac{dQ_A}{dt} = \frac{Q_{ava}v_s}{W_D}. \tag{50}$$

For maximum power efficiency, this current should drop near the end of the voltage cycle before the voltage rises above the mean value. Since the duration of the current pulse corresponds to the transit time of the charge packet and is equal to half cycle, the frequency of operation is optimized at

$$f = \frac{v_s}{2W_D}. \tag{51}$$

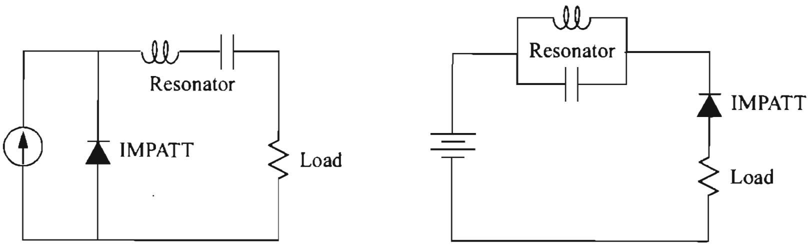
For practical oscillators the biasing circuits are shown in Fig. 9, and the current-source bias scheme is more common than the voltage-source. The external resonator circuit has a resonant frequency matching that of Eq. 51. Note that the terminal ac voltage across the IMPATT diode in the preceding discussions can be generated from a dc biasing circuit, the basic function of an oscillator. With a dc bias, any noise generated internally would be amplified because of positive feedback, until a stabilized ac waveform is established with the above current magnitude and frequency.

Figure 8 shows that positive ac current coincides with the negative ac voltage, a phase shift of ≈ π. This is the origin of dynamic negative resistance, or negative power absorbed by the device. Note that for the terminal current waveform, the start of the current pulse is determined by the injection delay, while the end is determined by the transit delay.

The capability of power generation from a transit-time device should not be confused with the phase difference also introduced in the ac characteristics of a capacitor or inductor. The phase difference in these passive devices between its terminal voltage and current is $\pi/2$. So for half of the cycle, power absorbed by the device is negative but is positive for the other half cycle. These cancel each other exactly and the net power is zero.

9.4.2 Power-Frequency Limitation—Electronic

Because of the inherent limitations of semiconductor materials and the attainable impedance levels in microwave circuitry, the maximum output power at a given fre-



(a) (b)

Fig. 9 Biasing circuits for IMPATT-diode oscillators, with (a) current source and (b) voltage source.

quency of a single diode is limited. The limitations on semiconductor materials are (1) the critical electric field $\mathscr{E}_m$ at which the avalanche breakdown occurs, and (2) the saturation velocity υ_s, which is the maximum attainable velocity in the semiconductor.

The maximum voltage that can be applied across a semiconductor sample is limited by the breakdown voltage, which, for a maximum, uniform avalanche region, is given by $V_m = \mathscr{E}_m W_D$. The maximum current that can be carried by the semiconductor sample is also limited by the avalanche-breakdown process, because the avalanched charge in the drift region causes a disturbance in the electric field. With the maximum avalanched charge Q_{ava} given by $\mathscr{E}_m \varepsilon_s$ (Gauss' law), the maximum current density is given by

$$J_m = \frac{\mathscr{E}_m \varepsilon_s \upsilon_s}{W_D} . \tag{52}$$

Therefore, the upper limit on the power density is given by the product of V_m and J_m:

$$P_m = V_m J_m = \mathscr{E}_m^2 \varepsilon_s \upsilon_s . \tag{53}$$

Combining with Eq. 51, Eq. 53 can be rewritten as

$$P_m f^2 \approx \frac{\mathscr{E}_m^2 \upsilon_s^2}{4\pi X_c} \tag{54}$$

where X_c is the reactance $(2\pi f C_D)^{-1}$. In practical high-speed oscillator circuits, X_c is found to be fixed due to interaction with some minimum external circuit impedance and the avalanche region that has been neglected.

Equation 54 thus predicts that the maximum power that an IMPATT diode can be designed for decreases as $1/f^2$. This electronic limit is expected to be dominant above the millimeter-wave frequencies (> 30 GHz) for Si and GaAs. For a practical operating junction temperature of 150° to 200°C, $\mathscr{E}_m$ in Si is about 10% smaller than that in GaAs. On the other hand, υ_s in Si is almost twice as large as in GaAs. Therefore, in the electronic-limited range (i.e., above millimeter-wave frequencies), the Si IMPATT diode is expected to have a power output about three times as large compared to that of GaAs, operated at the same frequency.[17] In the submillimeter-wave region, the uniform-field Misawa diodes are expected to be preferred because the device has a broad negative-resistance band and the transit-time effects do not play the dominating role in producing negative resistance as they do in Read diodes.[18] Under pulsed conditions where thermal effects can be ignored (i.e., short pulses), the peak power capability is determined by electronic limits (i.e., $P \propto 1/f^2$) at all frequencies.

9.4.3 Limitation on Efficiency

For efficient operation of an IMPATT diode, as carriers move through the drift region, as large a charge pulse Q_{ava} as possible must be generated in the avalanche region without reducing the electric field in the drift region below that required for velocity saturation. The motion of Q_{ava} through the drift region results in an ac voltage amplitude mV_D, where m is the modulation factor ($m \leq 1$) and V_D is the average voltage developed across the drift region. At the optimum frequency ($\approx v_s/2W_D$), the motion of Q_{ava} also results in an alternating particle current that has a phase delay of ϕ with the ac voltage across the diode. If the average of the particle current is J_0, the particle current swing is at most from zero to $2J_0$. For a square waveform of particle current and a sinusoidal variation of drift voltage, both with magnitude and phase as described above, the microwave power generating efficiency η is [19,20]

$$\eta \equiv \frac{\text{ac power output}}{\text{dc power input}} = \frac{(2J_0/\pi)(mV_D)}{J_0(V_A + V_D)}|\cos\phi|$$
$$= \left(\frac{2m}{\pi}\right)\frac{|\cos\phi|}{1 + (V_A/V_D)} , \tag{55}$$

where V_A and V_D are the dc voltage drops across the avalanche region and the drift region respectively and their sum is the total applied dc voltage. The angle ϕ is the injection phase delay of the particle current. Under ideal conditions, ϕ is π and $|\cos\phi| = 1$. For double-drift diodes, the voltage V_D is replaced by $2V_D$. The ac power contribution from the avalanche region is neglected because the avalanche-region voltage is inductively reactive relative to the particle current. The displacement current is capacitively reactive relative to the diode voltage and, therefore, contributes no average ac power.

Equation 55 clearly shows that to improve the efficiency one must increase the ac voltage modulation factor m, optimize the phase delay angle toward π, and reduce the V_A/V_D ratio. However, V_A must be sufficiently large to initiate the avalanche process rapidly; below a certain optimum value of V_A/V_D, the efficiency falls off toward zero.[19]

If the drifting carriers are velocity saturated at very low field, m could approach unity and no deleterious consequence would result. In n-type GaAs, the velocity is effectively saturated near $\mathscr{E} \approx 10^3$ V/cm, which is much smaller than the field value for n-type Si, $\approx 2\times10^4$ V/cm. Hence, much larger ac voltage swings can be expected in n-GaAs; these larger voltage swings, in turn, result in higher efficiency in n-GaAs.[21]

To estimate the optimum value of V_A/V_D, we first get

$$V_D = \langle \mathscr{E}_D \rangle (W_D - x_A) = \frac{\langle \mathscr{E}_D \rangle v_s}{2f} \tag{56}$$

where $\langle \mathscr{E}_D \rangle$ is the average field in the drift region. For 100% current modulation, $J_0 = J_{dc} = J_{ac}$, and a maximum charge $Q_{ava} = m\varepsilon_s \langle \mathscr{E}_D \rangle$ determines the current density:

$$J_0 = Q_{ava} f = m\varepsilon_s \langle \mathscr{E}_D \rangle f. \tag{57}$$

For an ionization coefficient having a field dependence of $\alpha \propto \mathscr{E}^\zeta$ where ζ is a constant, the value of α' can be obtained as

$$\alpha' \equiv \frac{d\alpha}{d\mathscr{E}} = \frac{\zeta\alpha}{\mathscr{E}} \approx \frac{\zeta(W_D - x_A)\alpha}{V_D} . \tag{58}$$

Assume that the transit-time frequency (Eq. 51) is about 20% larger than the resonant frequency (Eq. 39), combining Eqs. 56 through 58 yields [20]

$$\left.\frac{V_A}{V_D}\right|_{\text{opt}} \approx 4m\left(\frac{1.2}{2\pi}\right)^2 \zeta \alpha x_A . \tag{59}$$

For the relatively low frequency of ≈ 10 GHz, the optimum value of V_A/V_D for GaAs is 0.65 with $m = 1$, while for Si the optimum value is about 1.1 with $m = 1/2$.

A plot of the efficiency versus V_A/V_D is shown in Fig. 10. The maximum efficiency is obtained using the optimum values discussed above. The expected maximum efficiency is about 15% for single-drift (SD) Si diodes, 21% for double-drift (DD) Si diodes, and 38% for single-drift GaAs diodes. The foregoing estimates are consistent with experimental results. At higher frequencies, the optimum ratio V_A/V_D tends to increase; this increase results in a reduction of the maximum efficiency. The experimental results for *n*-GaAs single-drift diodes are also in agreement with the foregoing discussion.[22]

In practical IMPATT diodes, many other factors reduce the efficiency. These factors include the space-charge effect, reverse saturation current, series resistance, skin effect, saturation of ionization rate, tunneling, intrinsic avalanche response time, minority-carrier storage, and thermal effect.

The space-charge effect is shown in Fig. 11.[23] The generated electrons will depress the field (Fig. 11a). The reduction in field may turn off the avalanche process prematurely and reduce the 180° phase delay provided by the avalanche. As the electrons drift to the right (Fig. 11b), the space charge may also cause the field to the left of the carrier pulse to drop below that required for velocity saturation. This drop, in turn, will change the terminal current waveforms and reduce the power generated at the transit-time frequency.

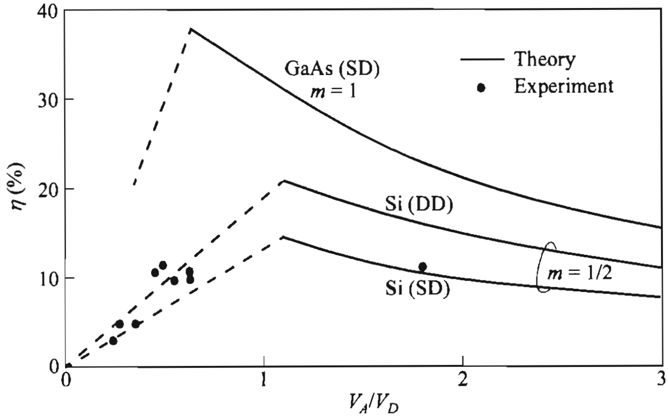


Fig. 10 Efficiency vs. ratio V_A/V_D for Si and GaAs diodes. SD, DD = single, double drift. Dashed lines are estimated linear extrapolation from peak to zero. (After Ref. 20.)

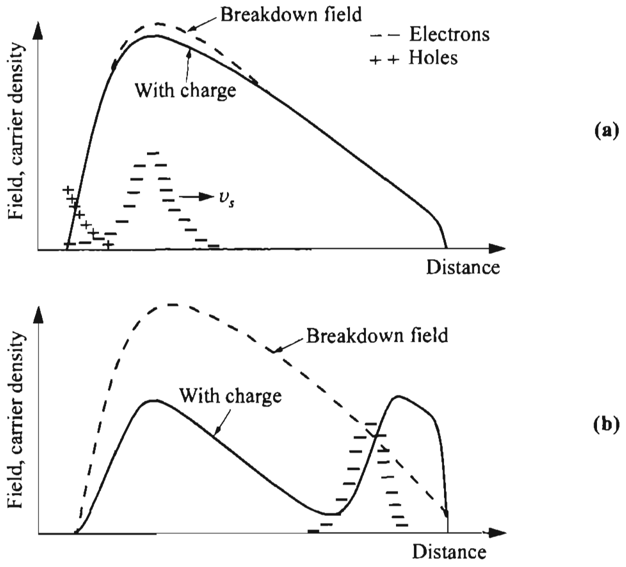


Fig. 11 Instantaneous field and charge distributions in a Read diode. (a) Avalanching just completed, charge beginning to move across diode. (b) Charge transit nearly completed. Note the strong effect of the space charge in depressing the electric field. (After Ref. 23.)

A high reversed saturation current causes the avalanche to build up too soon, reducing the avalanche phase delay and thus the efficiency.[24] The minority injection from a poor ohmic contact will also increase the reverse saturation current and thus reduce the efficiency.

Near the end of the drift region, the field is smaller. The carriers travel in the mobility regime at a reduced speed from the saturation velocity.[25] The unswept layer gives rise to a series resistance that reduces the terminal negative resistance. Note, however, that the effect on *n*-GaAs is much smaller, because GaAs has much higher low-field mobility.

As the operating frequency of an IMPATT diode is increased into the millimeter-wave region, the current will be confined to flow within a skin depth δ of the surface of the substrate. The skin effect is shown in Fig. 12.[26] Thus, the effective resistance of the substrate is increased, giving rise to a voltage drop across the radius of the diode (Fig. 12b). This voltage drop will cause nonuniform current distribution in the diode and a high effective series resistance, both of which reduce efficiency. However, advanced fabrication technologies have effectively eliminated this issue of skin effect, and it only plays a minor role in some upside-down mounted devices.

For very-high-frequency operation, the depletion width has to be quite narrow (Eq. 51) and the field required for impact ionization becomes higher in order to satisfy the integral criterion of Eq. 2. There are two major effects at such high fields. The first effect is that the ionization rate will vary slowly at high field, broadening the

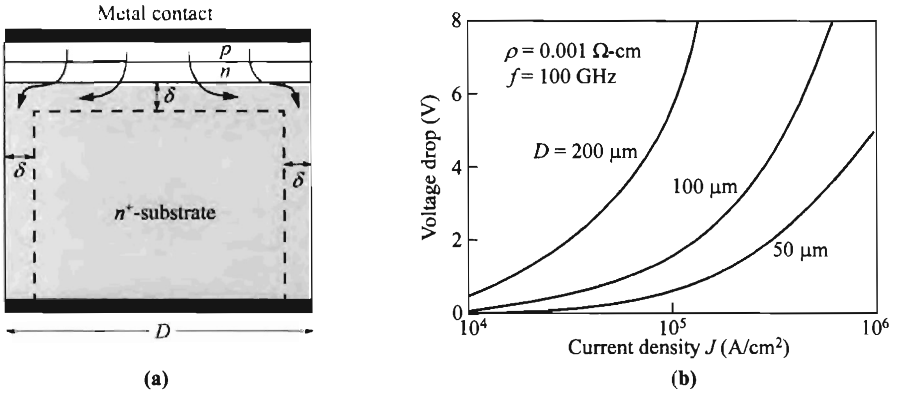


Fig. 12 Skin effect in IMPATT diode. (a) Current flow confined to a surface lamina of thickness δ, causing nonuniformity and resistive loss. (b) Calculated voltage drop in the substrate at 100 GHz for several diode diameters D. (After Ref. 26.)

injected current pulse[27] and changing the terminal current waveforms so that the efficiency is reduced. The second effect is the tunneling current, which may be dominant. Since it is in phase with the field, the 180° avalanche phase delay is absent.

Another factor that limits performance at submillimeter waves is the finite delay by which the ionization rate lags the electric field. For Si this *intrinsic avalanche response time* τ_i is less than 10^{-13} s. Since this time is very small compared to the transit time in the submillimeter-wave region, Si IMPATT diodes are expected to be efficient up to 300 GHz or higher frequencies. For GaAs, however, τ_i is found to be an order of magnitude longer than that of Si.[28] Such a long τ_i may limit the GaAs IMPATT operation to frequencies below 100 GHz.

Minority-carrier storage effects in p^+-n (or n^+-p) diodes arising from back diffusion of the generated electrons (or holes) from the active layer into the neutral p^+- (or n^+-) region can occur and will degrade the efficiency. This minority carrier will be stored in the neutral region while the remaining carriers are in transit and will diffuse back into the active region at a later time in the cycle, causing a premature avalanche which destroys the current-voltage phase relationship.

9.4.4 Power-Frequency Limitation—Thermal

At lower frequencies, the cw performance of an IMPATT diode is limited primarily by thermal considerations, that is, by the power that can be dissipated in a semiconductor chip. A typical device mounting arrangement is that the IMPATT diode is bonded upside down onto the substrate of good thermal conductivity, such that the heat source is closest to the heat sink. If the contacts to the top diode surface have multiple layers of metals, the total thermal resistance is given by the series combination of[29]

$$R_T = \sum \frac{d_s}{A\kappa_s} + \sum \frac{1}{\pi\kappa_h R_h}\left[1 + \frac{z_h}{R_h} - \sqrt{1 + \left(\frac{z_h}{R_h}\right)^2}\right], \tag{60}$$

where d_s and κ_s are each layer thickness and its thermal conductivity of metal layers on the diode surface, while z_h, κ_h, and R_h are each layer thickness, its thermal conductivity, and the contact radius (close to the device) of the heat sink. For a single-layered semi-infinite heat sink, z_h/R_h approaches infinity and the second term reduces to the familiar expression of $1/\pi\kappa_h R_h$. Copper and diamond are two common heat-sink materials. Since diamond has a thermal conductivity three times that of copper, there is a trade-off between performance and cost.

The power P which can be dissipated in the diode must equal the heat power that can be transmitted to the heat sink. Therefore, P equals $\Delta T/R_T$, where ΔT is the temperature difference between the junction and the heat sink. If the reactance $X_c = 2\pi f C_D$ is maintained constant ($f \propto 1/C_D$) and the major contribution to the thermal resistance is from the semiconductor (assuming that $d_s \approx W_D$, $R_T = W_D/A\kappa_s$), we obtain for a given temperature increase ΔT,

$$P \cdot f = \left(\frac{\Delta T}{R_T}\right) f \approx \frac{\Delta T}{W_D/A\kappa_s}\left(\frac{W_D}{A\varepsilon_s}\right) = \frac{\kappa_s \Delta T}{\varepsilon_s} = \text{constant}. \tag{61}$$

Under such conditions, the cw power output will decrease as $1/f$. Therefore, under cw conditions, at lower frequencies we have thermal limitation ($P \propto 1/f$) and at higher frequencies we have electronic limitation ($P \propto 1/f^2$). The corner frequency at which the power drops rapidly for a given semiconductor depends on the maximum allowed temperature rise, the minimum attainable circuit impedance, and the product of $\mathscr{E}_m$ and v_s.

Burnout from Filament Formation. Burnout may occur not only if the diode is overheated, but also, more insidiously, if the carrier current fails to be uniformly distributed over the diode area and is instead concentrated into filaments of locally high intensity. Such untoward behavior can often result when the diode has a dc negative conductance because, then, the local region of greatest current density also has the lowest breakdown voltage. For this reason, *p-i-n* diodes are prone to easy burnout. The moving carrier space charge in the drift region tends to prevent low-frequency negative resistance and therefore helps to prevent filamentary burnout. Diodes that have positive dc resistance at low currents may develop negative dc resistance and burnout at high currents.

9.5 NOISE BEHAVIOR

The noise in an IMPATT diode arises mainly from the statistical nature of the generation rates of electron-hole pairs in the avalanche region. Since the noise sets a lower limit to the microwave signals to be amplified, it is important to consider the noise theory of the IMPATT diode.

For amplification, the IMPATT diode can be inserted into a resonator that is coupled to a transmission line.[30] The line is coupled to separate input and output by means of a circulator, as shown in Fig. 13a. Figure 13b shows the equivalent circuit upon which the small-signal analysis is based. We shall now introduce two useful expressions for the noise performance: the *noise figure* and the *noise measure*. The noise figure NF is defined as

$$NF = 1 + \frac{\text{output noise power from amplifier}}{(\text{power gain}) \times (kT_0 B_1)} = 1 + \frac{\langle I_n^2 \rangle R_L}{G_P k T_0 B_1} \tag{62}$$

where G_P is the amplifier power gain, R_L is the load resistance, T_0 = room temperature (290 K), B_1 is the noise bandwidth, and $\langle I_n^2 \rangle$ is the mean-square noise current caused by the diode and induced in the loop of Fig. 13b. The noise measure M is defined as

$$M \equiv \frac{\langle I_n^2 \rangle}{4kT_0 G B_1} = \frac{\langle V_n^2 \rangle}{4kT_0(-Z_{\text{real}})B_1}, \tag{63}$$

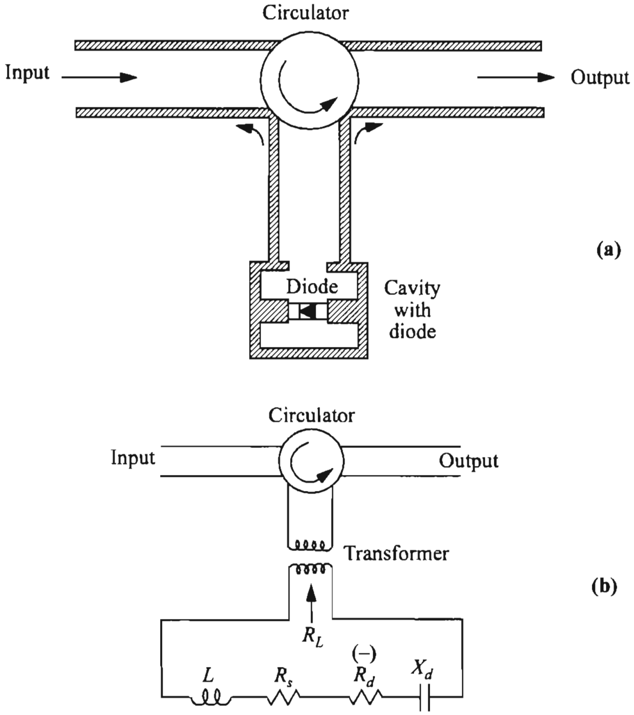


Fig. 13 (a) IMPATT diode inserted into a resonator. (b) Equivalent circuit. (After Ref. 30.)

where G is the negative conductance, $-Z_{\text{real}}$ the real part of the diode impedance, and $\langle V_n^2 \rangle$ the mean-square noise voltage. Note that both the noise figure and the noise measure depend on the mean-square noise current (or the mean-square noise voltage). It will be shown that for frequencies above the resonant frequency f_r, the noise in the diode decreases, but so does the negative resistance. In this situation the appropriate quantity for assessing the performance of the diode as an amplifier is the noise measure, whose minimum value (minimum noise measure) is of special interest.

The noise figure for a high-gain amplifier is given by[30]

$$NF = 1 + \frac{qV_A/kT_0}{4\zeta\tau_A^2(\omega^2 - \omega_r^2)} \tag{64}$$

where τ_A and V_A are, respectively, the time and voltage drop across the avalanche region; and ω_r is the resonant frequency given in Eq. 39. The expression above is obtained under the simplified assumptions that the avalanche region is narrow and that the ionization coefficients of holes and electrons are equal. For $\zeta = 6$ (for Si) and $V_A = 3$ V, the noise figure at $f = 10$ GHz ($\omega = \omega_r$) is predicted to be 11,000 or 40.5 dB.

With realistic ionization coefficients ($\alpha_n \neq \alpha_p$ for Si) and an arbitrary doping profile, the low-frequency expression for the mean-square noise voltage is given by[31]

$$\langle V_n^2 \rangle = \frac{2qB_1}{J_0 A}\left[\frac{1 + (W_D/x_A)}{\alpha'}\right]^2 \propto \frac{1}{J_0} \tag{65}$$

where $\alpha' = \partial\alpha/\partial\mathscr{E}$. Figure 14 shows $\langle V_n^2 \rangle / B_1$ as a function of frequency for a silicon IMPATT diode with $A = 10^{-4}$ cm^2, $W_D = 5$ μm, and $x_A = 1$ μm. At low frequencies, note that the noise voltage $\langle V_n^2 \rangle$ is inversely proportional to the direct current, Eq. 65. Near the resonant frequency (which varies as $\sqrt{J_0}$), $\langle V_n^2 \rangle$ reaches a maximum and then decreases roughly as the fourth power of frequency. Noise can therefore be reduced somewhat by operating well above the resonant frequency and keeping the

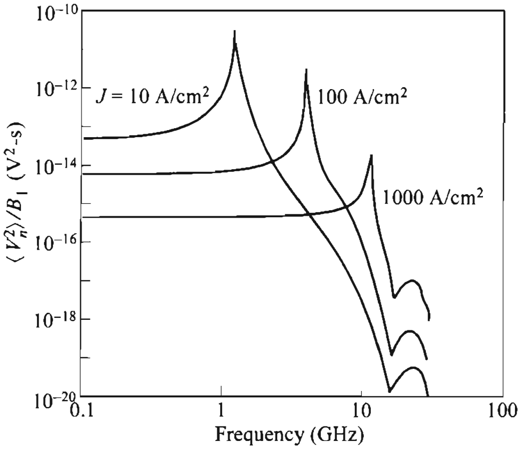


Fig. 14 Mean-square noise voltage over bandwidth vs. frequency for a Si IMPATT diode. (After Ref. 31.)

Table 1 Noise Measure of IMPATT Diodes

Semiconductor	Ge	Si	GaAs
Small-signal noise measure (dB)	30	40	25
Large-signal oscillator noise measure (dB)	40	55	35

current low. These conditions conflict with conditions favoring high power and efficiency, so that trade-offs are necessary to optimize for particular applications.

Figure 15 shows typical theoretical and experimental results of the noise measure in a GaAs IMPATT diode. At the transit-time frequency (6 GHz), the noise measure is about 32 dB. The minimum noise measure of 22 dB, however, is obtained at about twice the transit-time frequency. One important feature of the GaAs noise measure is that it is substantially lower than that for Si IMPATT diodes. Table 1 compares the noise measures of Ge, Si, and GaAs IMPATT diodes. The amplifier and oscillator noises in the table are for a lossless circuit at a frequency that corresponds to maximum oscillator efficiency without harmonic tuning. More-recent result gives a lower noise measure of 22 dB at 60 GHz.[33]

The main reason for the low-noise behavior in GaAs is that for a given field the electron and hole ionization rates are essentially the same, whereas in Si they are quite different. From the avalanche multiplication integral, it can be shown that to obtain a large multiplication factor M the average distance of ionization $1/\langle\alpha\rangle$ is about equal to x_A (the avalanche width) if $\alpha_n = \alpha_p$, but is about $x_A/\ln(M)$ if $\alpha_n \gg \alpha_p$. So, for a given x_A, considerably more ionization events must occur in Si, resulting in higher noise.

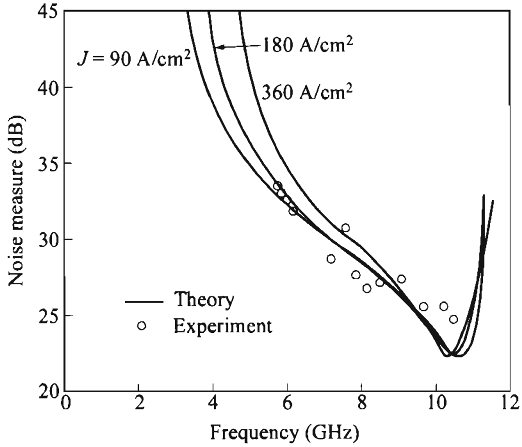


Fig. 15 Noise measure for GaAs IMPATT diodes. Transit-time frequency = 6 GHz. (After Ref. 32.)

Figure 16 shows the relation between power output and noise measure for some Si and GaAs 6-GHz IMPATT diodes.[34] The power level is expressed with respect to a reference power of 1 mW, that is, the power is given by $10 \log(P \times 10^3)$ dBm, where P is in watts. The diodes were evaluated in a single-tuned coaxial resonator circuit in which the load resistance presented to the resonator was incrementally varied by using interchangeable impedance transformers. At a maximum power output, the noise measure is relatively poor. A lower noise measure can be realized at the expense of a slightly reduced power output. Note again that at a given power level (say 1 W or 30 dBm), the GaAs IMPATT diode is about 10 dB quieter than a Si IMPATT diode.

9.6 DEVICE DESIGN AND PERFORMANCE

From the small-signal theory, we can obtain approximate relations for various device parameters as a function of operating frequency. Ignoring the small avalanche region x_A, the resistance expression in Eq. 47 can be rewritten as

$$-R \approx \frac{W_D^2}{2A\varepsilon_s v_s}\left[\frac{1}{(\omega^2/\omega_r^2)-1}\right]\left(\frac{1-\cos\theta_d}{\theta_d^2/2}\right) \tag{66}$$

where θ_d is the transit angle equal to $\omega W_D/v_s$. For fixed ω/ω_r, for $-R$ to be invariant it is required from Eq. 66 that both W_D^2/A (prefactor) and θ_d be constant. Since the depletion width W_D is inversely proportional to the operating frequency (Eq. 51), the device area A, which is proportional to W_D^2, is thus proportional to ω^{-2}. Also, from the avalanche breakdown equation, Eq. 2, it can be shown that the ionization rate (α) and its field derivative (α') are inversely proportional to the depletion-layer width

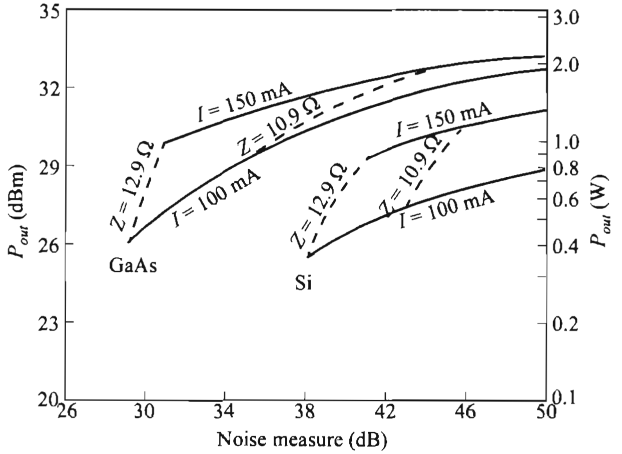


Fig. 16 Power output vs. noise measure for a phase-locked oscillator. Locking power was held constant at 4 dBm. Contours of constant load impedance Z and constant diode current I are shown. (After Ref. 34.)

W_D. Combining the relation $\alpha' \propto 1/W_D$ with Eq. 39 yields the following result for the dc current density:

$$J_0 \propto \frac{\omega_r^2}{\alpha'} \propto \frac{\omega^2}{1/W_D} \propto \omega. \tag{67}$$

These frequency scaling relations are summarized in Table 2 and are useful as a guide for extrapolating performance and design to new frequencies.

The power-output limitations have been considered in Section 9.4. The efficiency is expected to be only weakly dependent on frequency at low frequencies. However, at millimeter-wave regions the operating current density is high ($\propto f$) and the area is small ($\propto f^{-2}$), so that the device operating temperature will be high. This high temperature, in turn, will cause the reverse-saturation current to increase and the efficiency to decrease. In addition, the skin effect, tunneling, and other effects associated with high frequency and high field will also degrade the efficiency performance. Hence, as the frequency increases, the efficiency is expected to decrease eventually.

Figure 17 shows the dependence of the threshold current density, that is, the minimum current density to produce oscillation, on the frequency. Note that the threshold current density increases approximately as the square of the frequency, in agreement with the general behavior of the resonant frequency. To demonstrate the importance of the transit-time effect, Fig. 18 shows the optimum depletion-layer width versus the frequency for Si and GaAs IMPATT diodes. The depletion-layer width, as expected, varies inversely with the frequency. Interestingly, at frequencies above 100 GHz, the depletion-layer width is less than 0.5 μm. This very narrow width gives some indication of the difficulty inherent in fabricating a modified Read diode or a double-drift diode at these high frequencies.

The highest power·f^2 product is obtained from double-drift diodes. Figure 19 compares the performance of double-drift and single-drift diodes at 50 GHz. The double-drift 50-GHz Si IMPATT diode made by ion implantation shows an output cw power over 1 W with a maximum efficiency of 14%. This result can be compared with a similar single-drift diode that delivers about 0.5 W with an efficiency of 10%. The superiority of the double-drift diodes results from the fact that both holes and electrons produced by the avalanche are allowed to do work against the radio-fre-

Table 2 Frequency Scaling (Approximate) for IMPATT Diodes

Parameters		Frequency dependence
Junction area A		f^{-2}
Bias-current density J_0		f
Depletion-layer width W_D		f^{-1}
Breakdown voltage V_B		f^{-1}
Power output P_{out}:	Thermal limitation	f^{-1}
	Electronic limitation	f^{-2}
Efficiency η		Constant

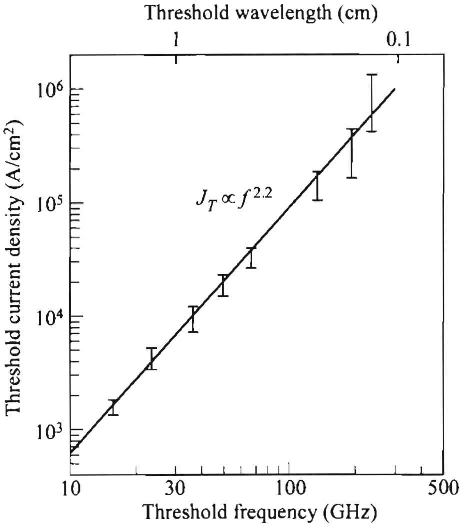


Fig. 17 Dependence of threshold frequency on direct current density. (After Ref. 35.)

quency (RF) field by traversing the drift regions. In the single-drift diodes only one type of carrier is so utilized. As a result, a larger terminal voltage can be applied.

A summary of the state-of-the-art IMPATT performance is given in Fig. 20. Also shown are results for BARITT diodes to be discussed in Section 9.7. At lower frequencies, the power output is thermal-limited and varies as f^{-1}; at higher frequencies (> 50 GHz) the power is electronic-limited and varies as f^{-2}. GaAs IMPATT diodes typically show better power performance at low frequencies below ≈ 60 GHz. Figure 20 clearly shows that the IMPATT diode is one of the most powerful solid-state sources for the generation of microwaves. The IMPATT diodes can generate higher cw power output in the millimeter-wave frequencies than any other solid-state

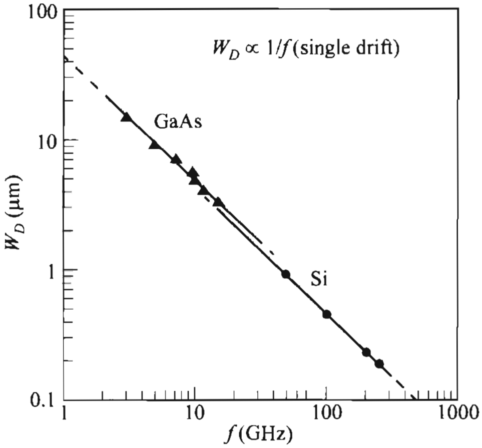


Fig. 18 Depletion width vs. frequency for Si and GaAs IMPATT diodes. (After Refs. 36 and 37.)

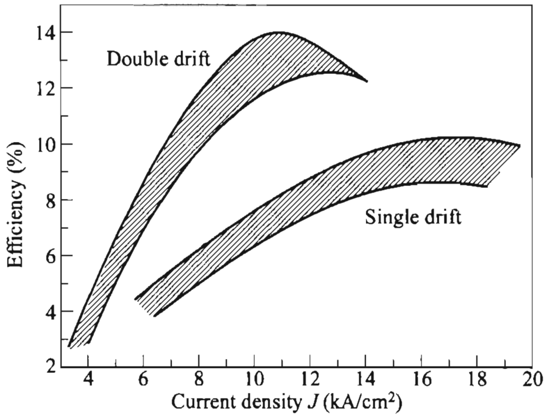


Fig. 19 Efficiency of single-drift vs. double-drift Si IMPATT diodes at 50 GHz. Range of efficiency for four diodes of each type. (After Ref. 38.)

device. For pulsed operations, the power can even be higher than that shown in Fig. 20.

Recently, materials other than Si and GaAs have been examined. For example, compared to Si, SiC has a ten times higher breakdown field, a thermal conductivity of three times, and a saturation velocity of two times.[40] These factors contribute to an expected power output of more than 350 times higher than Si. The disadvantage is a

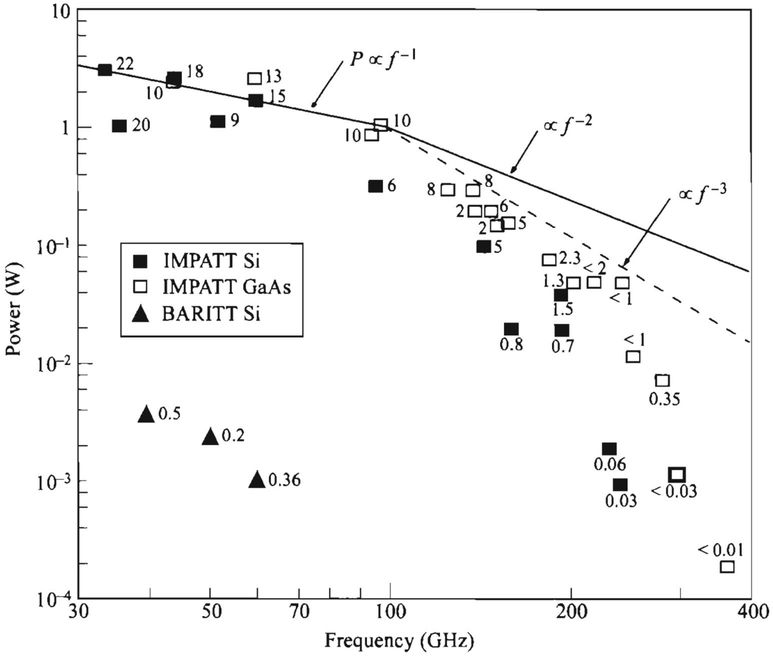


Fig. 20 State-of-the-art IMPATT (and BARITT) performances. The number against each experimental point indicates the efficiency in percentage. (After Ref. 39.)

higher noise measure. High-bandgap GaN also offers similar advantages, as well as operation at higher temperature.[41] In terms of structure, by incorporating a heterojunction at the injecting junction, GaN is expected to yield reduced leakage current, improved RF efficiency, and lower noise.[42]

9.7 BARITT DIODE

Another transit-time device is the BARITT (*barrier-injection transit-time*) diode. The mechanisms responsible for the microwave oscillation are the thermionic injection and diffusion of minority carriers across the forward-biased barrier. Because there is no avalanche delay time, the BARITT diode is expected to operate at lower efficiency and lower power than the IMPATT diode. On the other hand, the noise associated with carrier injection across the barrier is smaller than the avalanche noise in an IMPATT diode. The low-noise property and the stability of the device make the BARITT diode suited for low-power applications, such as local oscillators. The BARITT operation was first reported by Coleman and Sze in 1971 using a metal-semiconductor-metal reach-through diode.[43] Similar structures were proposed in 1968 by Ruegg based on large-signal analysis and by Wright using space-charge-limited transport mechanisms.[44,45]

9.7.1 Current Transport

The BARITT diode is basically a back-to-back pair of diodes biased into reach-through condition (Fig. 21). The two diodes can be *p-n* junctions or metal-semiconductor contacts, or combination of the two. We consider first the current transport in a symmetrical metal-semiconductor-metal (MSM) structure[46] with uniformly doped *n*-type semiconductor (Fig. 21b). With bias, the depletion-layer widths are

$$W_{D1} = \sqrt{\frac{2\varepsilon_s}{qN_D}(\psi_{bi} - V_1)}, \tag{68a}$$

$$W_{D2} = \sqrt{\frac{2\varepsilon_s}{qN_D}(\psi_{bi} + V_2)}, \tag{68b}$$

where W_{D1} and W_{D2} are the depletion widths in the *n*-layer for the forward- and reverse-biased barriers, respectively; V_1 and V_2 are the fraction of the applied voltage developed across the respective junctions, N_D is the ionized impurity density; and ψ_{bi} is the built-in potential. Under these conditions, the current is the sum of the reverse saturation current (of a Schottky diode with a barrier height ϕ_{Bn}), generation-recombination current, and surface leakage current.

As the voltage increases, the reverse-biased depletion region will eventually reach through to the forward-biased depletion region (Fig. 21c). The corresponding voltage is called the reach-through voltage V_{RT}. This voltage can be obtained from the condition $W_{D1} + W_{D2} = W$, which is the length of the *n*-region:

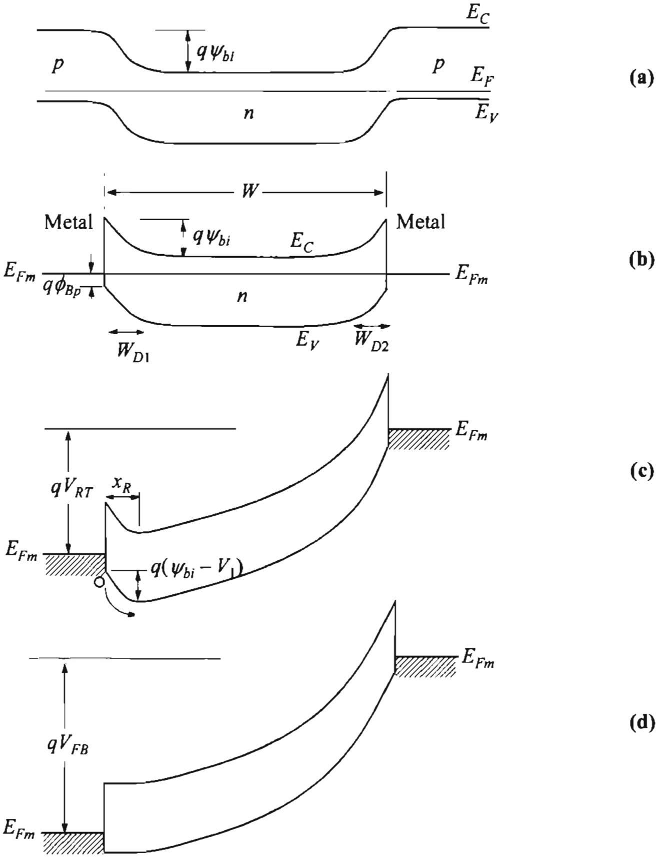

E_C
$q\psi_{bi}$
p
p
E_F
(a)
n
E_V
W
Metal
$q\psi_{bi}$
E_C
Metal
E_{Fm}
E_{Fm}
(b)
$q\phi_{Bp}$
n
E_V
W_{D2}
W_{D1}
E_{Fm}
qV_{RT}
x_R
(c)
E_{Fm}
$q(\psi_{bi}-V_1)$
E_{Fm}
(d)
qV_{FB}
E_{Fm}

Fig. 21 Energy-band diagrams of BARITT diodes having (a) *p-n* junctions and (b) M-S contacts, under thermal equilibrium. MSM structure at (c) punch-through and (d) flat-band conditions.

$$V_{RT} = \frac{qN_D W^2}{2\varepsilon_s} - W\sqrt{\frac{2qN_D}{\varepsilon_s}(\psi_{bi}-V_1)}$$
$$\approx \frac{qN_D W^2}{2\varepsilon_s} - W\sqrt{\frac{2qN_D\psi_{bi}}{\varepsilon_s}} \quad . \tag{69}$$

If the voltage is increased further, the energy band at the positively biased contact (left) can become flat. The electric field is zero at $x = 0$ when $\psi_{bi} = V_1$; this condition is the flat-band condition (Fig. 21d). The corresponding voltage is defined as the flat-band voltage V_{FB}:

$$V_{FB} \equiv \frac{qN_D W^2}{2\varepsilon_s} . \tag{70}$$

For a given length and with high doping levels, the applied voltage is limited by the avalanche breakdown voltage before reaching V_{FB}.

The dc bias for a BARITT diode under microwave oscillation is generally between V_{RT} and V_{FB}. For applied voltage in this range ($V_{RT} < V < V_{FB}$), the relation between the applied voltage and the forward-biased barrier height is

$$\psi_{bi} - V_1 = \frac{(V_{FB} - V)^2}{4V_{FB}} . \tag{71}$$

The reach-through point x_R as shown in Fig. 21c is given by

$$\frac{x_R}{W} = \frac{V_{FB} - V}{2V_{FB}} . \tag{72}$$

After reach-through, the hole current of thermionic emission over the hole barrier ϕ_{Bp} becomes the dominant current:

$$J_p = A_p^* T^2 \exp\left[-\frac{q(\phi_{Bp} + \psi_{bi})}{kT}\right]\left[\exp\left(\frac{qV_1}{kT}\right) - 1\right], \tag{73}$$

where A_p^* is the effective Richardson constant for holes (refer to Chapter 3). From Eq. 71 we obtain for $V \geq V_{RT}$,

$$J_p = A_p^* T^2 \exp\left(-\frac{q\phi_{Bp}}{kT}\right) \exp\left[-\frac{q(V_{FB} - V)^2}{4kTV_{FB}}\right] . \tag{74}$$

Therefore, beyond reach-through the current will increase exponentially with applied voltage.

At current levels high enough for the injected carrier density to be comparable to the background ionized-impurity density, the mobile carriers will influence the field distribution in the drift region. This is the space-charge effect. If all the mobile holes traverse the n-region with the saturation velocity v_s, and if $J > qv_sN_D$, the Poisson equation becomes

$$\frac{d\mathscr{E}}{dx} = \frac{\rho}{\varepsilon_s} = \frac{q}{\varepsilon_s}\left(N_D + \frac{J}{qv_s}\right) \approx \frac{J}{\varepsilon_s v_s} . \tag{75}$$

Integrating twice (with boundary conditions $\mathscr{E} = 0$, $V = 0$, at $x = 0$) yields[47]

$$J = \left(\frac{2\varepsilon_s v_s}{W^2}\right) V = qv_sN_D\left(\frac{V}{V_{FB}}\right) . \tag{76}$$

The current transport mechanisms of the reach-through p^+-n-p^+ structure is similar to that of the MSM structure. The only difference is that in Eqs. 73 and 74, the factor $\exp(-q\phi_{BP}/kT)$ should be absent when the carriers are injected over a forward-biased p^+-n junction,[47] that is,

$$J = A^* T^2 \exp\left[-\frac{q(V_{FB} - V)^2}{4kTV_{FB}}\right] . \tag{77}$$

For a PtSi-Si barrier, the hole barrier height $q\phi_{Bp}$ is equal to 0.2 eV. Hence at 300 K for a given voltage above reach-through, the current for the p^+-n-p^+ device will be about 3000 times larger than that for the MSM device. The value of A^*T^2 at room temperature is about 10^7 A/cm^2. Therefore, under normal operation, the onset of the space-charge effect will occur long before the flat-band condition.

A typical I-V characteristic is shown in Fig. 22 for a Si p^+-n-p^+ with a background doping of 5×10^{14} cm^{-3} and a thickness of 8.5 μm. The flat-band voltage is 29 V and reach-through voltage is about 21 V. We note that the current first increases exponentially and then becomes linearly dependent on voltage. The experimental results and theoretical calculation from Eqs. 76 and 77 are in good agreement.

For efficient BARITT operation, the current must increase very rapidly with voltage. The linear I-V relationship due to the space-charge effect will degrade device performance. The optimum current density is usually substantially lower than $J = qv_sN_D$.

9.7.2 Small-Signal Behaviors

We will show that the BARITT diode has small-signal negative resistance; therefore, the diode can have self-starting oscillation. Consider a p^+-n-p^+ structure. When it is biased above the reach-through voltage, the electric-field profile is shown in Fig. 23a. The point x_R corresponds to the potential maximum for hole injection, given by Eq. 72. The point a separates the low-field region from the saturation-velocity region, that is, for $\mathscr{E} > \mathscr{E}_s$, $v = v_s$, as shown in Fig. 23b. Under low-injection conditions,

$$a \approx \frac{\varepsilon_s \mathscr{E}_s}{qN_D} + x_R . \tag{78}$$

The transit time in the drift region ($x_R < x < W$) is given by:[49]

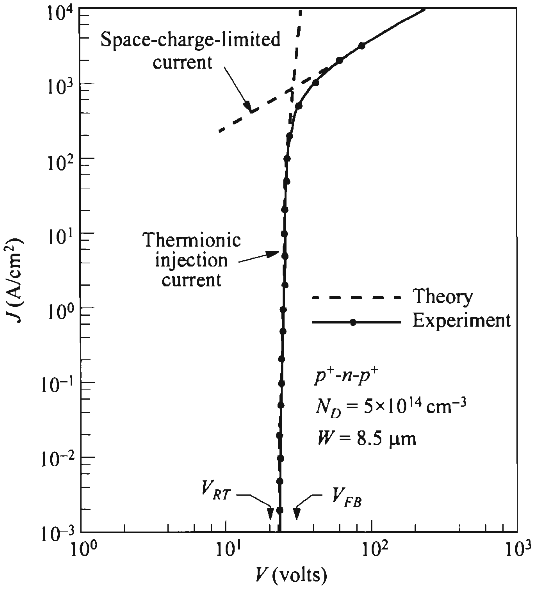


Fig. 22 Current-voltage characteristics of a Si p^+-n-p^+ reach-through diode. (After Ref. 47.)

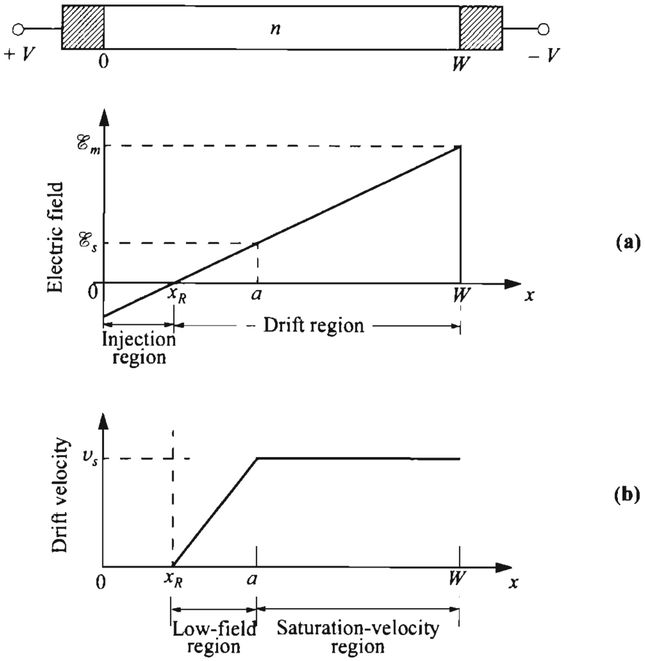

n
+V
0
W
−V
$\mathscr{E}_m$
Electric field
$\mathscr{E}_s$
0
x_R
a
W
x
Injection region
Drift region
(a)
Drift velocity
v_s
0
x_R
a
W
x
Low-field region
Saturation-velocity region
(b)

Fig. 23 (a) Field distribution and (b) carrier drift velocity in the drift region of a BARITT diode. (After Ref. 48.)

$$\tau_d = \int_{x_R}^{a} \frac{dx}{\mu_n \mathscr{E}(x)} + \int_a^W \frac{dx}{v_s} = \int_{x_R}^{a} \frac{dx}{\mu_n q N_D x/\varepsilon_s} + \frac{W-a}{v_s}$$
$$\approx \frac{3.75\varepsilon_s}{q\mu_n N_D} + \frac{W-a}{v_s} \quad . \tag{79}$$

To derive the small-signal impedance, we shall follow an approach similar to that used in Section 9.3.2 and introduce the time-varying quantity as the sum of a time-independent term (dc component) and a small ac term:

$$J(t) = J_0 + \tilde{J}\exp(j\omega t), \tag{80}$$

$$V(t) = V_0 + W\tilde{\mathscr{E}}\exp(j\omega t). \tag{81}$$

Substituting the foregoing expressions into Eq. 77 yields the linearized ac injected hole current density:

$$\tilde{J} = \sigma\tilde{\mathscr{E}} \tag{82}$$

where σ is the injection conductance per unit area and is given by

$$\sigma = J_0 \frac{\varepsilon_s(V_{FB} - V_0)}{N_D W k T} . \tag{83}$$

J_0 is the current density given by Eq. 77, in which V is replaced by V_0. The injection conductance increases with the applied voltage, reaches a maximum, and then decreases rapidly when V_0 approaches V_{FB}. The bias voltage corresponding to the maximum σ can be derived from Eqs. 77 and 83:

$$V_0(\text{for max } \sigma) = V_{FB} - \sqrt{\frac{2kTV_{FB}}{q}} . \tag{84}$$

Since the ac electric field is continuous across the boundary of the injection and the drift regions, these two regions will interact with each other. We define $\tilde{J}$ as the total alternating current density and $\tilde{J}_1$ as the injection current density. We assume that the injection region is thin enough that $\tilde{J}_1$ enters the drift region without delay. The alternating conduction current density in the drift region is then given by

$$\tilde{J}_c(x) = \tilde{J}_1 \exp[-j\omega\tau(x)] \equiv \gamma\tilde{J}\exp[-j\omega\tau(x)] \tag{85}$$

which is an unattenuated wave propagating toward $x = W$ with a transit phase delay of $\omega\tau(x)$. The quantity $\gamma \equiv \tilde{J}_1/\tilde{J}$ is the complex fraction relating the ac injection current to the total alternating current.

At a given position in the drift region, the total alternating current $\tilde{J}$ is equal to the sum of the conduction current $\tilde{J}_c$ and the displacement current $\tilde{J}_d$:

$$\tilde{J} = \tilde{J}_c(x) + \tilde{J}_d(x) \tag{86}$$

which is constant and independent of x. The displacement current is related to the ac field $\tilde{\mathscr{E}}(x)$ by

$$\tilde{J}_d(x) = j\omega\varepsilon_s \tilde{\mathscr{E}}(x) . \tag{87}$$

Combining Eqs. 83, 85, and 87 yields an expression for the ac electric field in the drift regions as a function of x and $\tilde{J}$,

$$\tilde{\mathscr{E}}(x) = \frac{\tilde{J}}{j\omega\varepsilon_s}\{1 - \gamma\exp[-j\omega\tau(x)]\} . \tag{88}$$

Integrating $\tilde{\mathscr{E}}(x)$ gives the ac voltage across the drift region in terms of the ac current density $\tilde{J}$. The coefficient can be expressed as

$$\gamma = \frac{\tilde{J}_1}{\tilde{J}_1 + \tilde{J}_d} = \frac{\sigma}{\sigma + j\omega\varepsilon_s} . \tag{89}$$

Substituting γ into Eq. 88 and integrating over the drift length $(W - x_R)$ with the boundary conditions of $\tau = 0$ at $x = x_R$ and $\tau = \tau_d$ at $x = W$ yields the expression for the ac voltage across the drift region:

$$V_d = \frac{\tilde{J}(W - x_R)}{j\omega\varepsilon_s}\left[1 - \left(\frac{\sigma}{\sigma + j\omega\varepsilon_s}\right)\frac{1 - \exp(j\theta_d)}{j\theta_1}\right] \tag{90}$$

where θ_d is the transit angle in the drift region,

$$\theta_d = \omega\left(\frac{W-a}{v_s} + \frac{3.75\varepsilon_s}{q\mu_n N_D}\right) = \omega\tau_d \tag{91}$$

and θ_1 is a constant given by

$$\theta_1 \equiv \omega\left(\frac{W-x_R}{v_s}\right). \tag{92}$$

We can also define $C_D = \varepsilon_s/(W-x_R)$ as the capacitance of the drift region. From Eq. 90 we obtain the small-signal impedance of the structure

$$Z \equiv \frac{\tilde{V}_d}{\tilde{J}} = R_d - jX_d \tag{93}$$

where R_d and X_d are the small-signal resistance and reactance, respectively:

$$R_d = \frac{1}{\omega C_D}\left(\frac{\sigma}{\sigma^2 + \omega^2\varepsilon_s^2}\right)\left[\frac{\sigma(1-\cos\theta_d) + \omega\varepsilon_s\sin\theta_d}{\theta_1}\right], \tag{94}$$

$$X_d = \frac{1}{\omega C_D} - \frac{1}{\omega C_D}\left(\frac{\sigma}{\sigma^2 + \omega^2\varepsilon_s^2}\right)\left[\frac{\sigma\sin\theta_d - \omega\varepsilon_s(1-\cos\theta_d)}{\theta_1}\right]. \tag{95}$$

Note that the real part (resistance) will be negative if the transit angle θ_d lies between the values of π and 2π, and if $|(1-\cos\theta_d)/\sin\theta_d|$ is less than $\omega\varepsilon_s/\sigma$.

From these results, we have shown that (1) the BARITT diodes have negative small-signal resistances and therefore can have self-starting oscillation; (2) the injection over the forward-biased p^+-n junction or metal-semiconductor barrier can serve as the source of carriers; and (3) the transit time in the drift region is important for the frequency characteristics of the BARITT diodes.

The BARITT diode has been shown to be a low-noise device, with basically only two noise sources. One noise source is the shot noise of the injected carriers (injection noise). The other noise source is the random velocity fluctuation of the carriers (diffusion noise) in the drift region.

9.7.3 Large-Signal Performance

The basic large-signal BARITT operation is shown in Fig. 24.[50] The carriers are injected as a delta function when the ac voltage reaches its peak ($\theta = \pi/2$). The induced external current travels three-fourths of a cycle to reach the negative terminal:

$$\theta_d = \omega\tau_d = \frac{3\pi}{2} \tag{96}$$

or

$$f = \frac{3}{4\tau_d} \approx \frac{3v_s}{4W}, \tag{97}$$

where θ_d is the transit angle and τ_d the carrier transit time. More accurate values for the optimum frequency can be obtained by substituting Eq. 79 into Eq. 97.

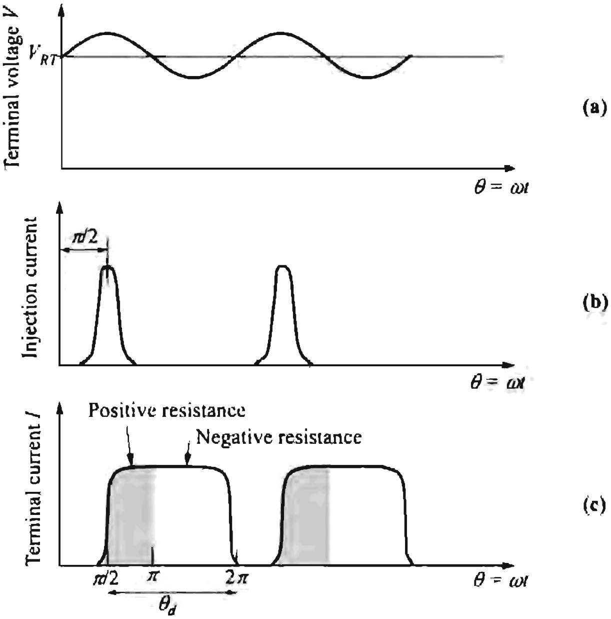


Fig. 24 (a) Terminal-voltage, (b) injection-current, and (c) terminal-current waveforms for a BARITT diode.

The maximum efficiency of BARITT diodes has been estimated to be of the order of 10% if the carriers are injected at $\theta = \pi/2$ (Fig. 24). However, higher efficiencies can be obtained if the carrier injection can be further delayed, that is, $\pi/2 < \theta \leq \pi$. A multilayered n^+-i-p-ν-n^+ BARITT diode,[51] which is a complementary structure of p^+-i-n-π-p^+, has been fabricated. The n^+-i-p region serves as a retarding field to increase the injection delay time. The state-of-the-art BARITT performance is shown in Fig. 20. Although the power output near 50 GHz is about two orders of magnitude smaller than that of IMPATT diodes, so is the noise measure. By optimizing the injection delay processes, the BARITT diode is expected to realize its full potential as a low-noise microwave source with moderate power and efficiency.

9.8 TUNNETT DIODE

For very-high-frequency operation, the depletion width in the IMPATT diode becomes quite narrow and the field required for impact ionization becomes high. There are two major effects at such high fields. The first is that the ionization rate will vary slowly at high field, broadening the injected current pulse and changing the terminal current waveforms so that the efficiency is reduced.[27] The second effect is the tunneling current which may be dominant. Since this tunneling current is in phase with the field, the 180° avalanche phase delay is not provided. The tunneling mecha-

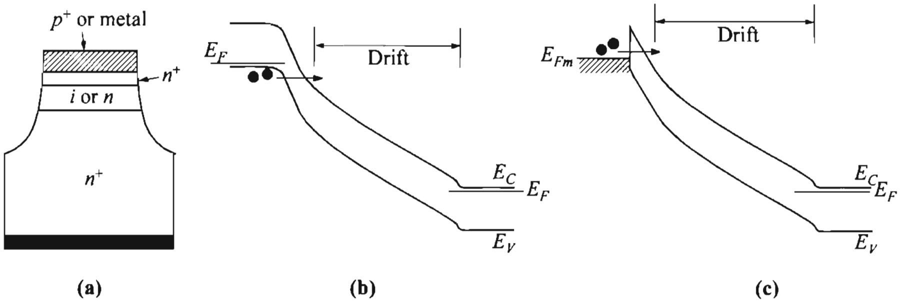


Fig. 25 (a) Structure of a TUNNETT diode. Energy-band diagrams showing (b) band-to-band tunneling in a *p-n* junction injector and (c) tunneling through the barrier in a Schottky-barrier injector.

nism has been considered for the TUNNETT (tunnel-injection transit-time) mode of operation.[2,52] The TUNNETT diode is expected to have lower noise than the IMPATT diode; however, the power output and efficiency will also be much lower.

In the TUNNETT diode, the tunneling injection current occurs at a high field of ≈ 1 MV/cm. The structure is different from a BARITT diode in that only one junction exists. The vicinity of the injecting junction has a higher doping level (Fig. 25). A typical n^+-layer (for an n-type drift region) near the injector has a doping of $\approx 10^{19}$ cm^{-3} and a thickness of ≈ 10 nm. Tunneling can occur band-to-band in the case of a *p-n* junction injector, and through the barrier in the case of a Schottky barrier. The advantages of a TUNNETT diode include the high-frequency capability that can go up to 1 THz theoretically. Continuous-wave generation higher than 650 GHz has been observed.[53] Another advantage is the low-voltage operation, which can be as low as 2 V. The limitation on the device is the power capability due to low tunneling current.

REFERENCES

1. W. Shockley, "Negative Resistance Arising from Transit Time in Semiconductor Diodes," *Bell Syst. Tech. J.*, **33**, 799 (1954).
2. W. T. Read, "A Proposed High-Frequency Negative Resistance Diode," *Bell Syst. Tech. J.*, **37**, 401 (1958).
3. R. L. Johnston, B. C. DeLoach, Jr., and B. G. Cohen, "A Silicon Diode Oscillator," *Bell Syst. Tech. J.*, **44**, 369 (1965).
4. B. C. DeLoach, Jr., "The IMPATT Story," *IEEE Trans. Electron Dev.*, **ED-23**, 57 (1976).
5. C. A. Lee, R. L. Batdorf, W. Wiegman, and G. Kaminsky, "The Read Diode and Avalanche, Transit-Time, Negative-Resistance Oscillator," *Appl. Phys. Lett.*, **6**, 89 (1965).
6. T. Misawa, "Negative Resistance on *p-n* Junction under Avalanche Breakdown Conditions, Parts I and II," *IEEE Trans. Electron Dev.*, **ED-13**, 137 (1966).

7. M. Gilden and M. F. Hines, "Electronic Tuning Effects in the Read Microwave Avalanche Diode," *IEEE Trans. Electron Dev.*, **ED-13**, 169 (1966).
8. C. A. Brackett, "The Elimination of Tuning Induced Burnout and Bias Circuit Oscillation in IMPATT Oscillators," *Bell Syst. Tech. J.*, **52**, 271 (1973).
9. G. Salmer, H. Pribetich, A. Farrayre, and B. Kramer, "Theoretical and Experimental Study of GaAs IMPATT Oscillator Efficiency," *J. Appl. Phys.*, **44**, 314 (1973).
10. S. M. Sze and G. Gibbons, "Avalanche Breakdown Voltages of Abrupt and Linearly Graded *p-n* Junctions in Ge, Si, GaAs, and GaP," *Appl. Phys. Lett.*, **8**, 111 (1966).
11. W. E. Schroeder and G. I Haddad, "Avalanche Region Width in Various Structures of IMPATT Diodes," *Proc. IEEE*, **59**, 1245 (1971).
12. G. Gibbons and S. M. Sze, "Avalanche Breakdown in Read and *p-i-n* Diodes," *Solid-State Electron.*, **11**, 225 (1968).
13. C. R. Crowell and S. M. Sze, "Temperature Dependence of Avalanche Multiplication in Semiconductors," *Appl. Phys. Lett.*, **9**, 242 (1966).
14. H. C. Bowers, "Space-Charge-Limited Negative Resistance in Avalanche Diodes," *IEEE Trans. Electron Dev.*, **ED-15**, 343 (1968).
15. S. M. Sze and W. Shockley, "Unit-Cube Expression for Space-Charge Resistance," *Bell Syst. Tech. J.*, **46**, 837 (1967).
16. P. Weissglas, "Avalanche and Barrier Injection Devices" in M. J. Howes and D. V. Morgan, Eds., *Microwave Devices–Device Circuit Interactions*, Wiley, New York, 1976, Chap. 3.
17. D. L. Scharfetter, "Power-Impedance-Frequency Limitation of IMPATT Oscillators Calculated from a Scaling Approximation," *IEEE Trans. Electron Dev.*, **ED-18**, 536 (1971).
18. H. W. Thim and H. W. Poetze, "Search for Higher Frequencies in Microwave Semiconductor Devices," 6th Eur. Solid State Device Res. Conf., *Inst. Phys. Conf. Ser.*, **32**, 73 (1977).
19. D. L. Scharfetter and H. K. Gummel, "Large-Signal Analysis of a Silicon Read Diode Oscillator," *IEEE Trans. Electron Dev.*, **ED-16**, 64, (1969).
20. T. E. Seidel, W. C. Niehaus, and D. E. Iglesias, "Double-Drift Silicon IMPATTs at X Band," *IEEE Trans. Electron Dev.*, **ED-21**, 523 (1974).
21. P. A. Blakey, B. Culshaw, and R. A. Giblin, "Comprehensive Models for the Analysis of High Efficiency GaAs IMPATTs," *IEEE Trans. Electron Dev.*, **ED-25**, 674 (1978).
22. K. Nishitani, H. Sawano, O. Ishihara, T. Ishii, and S. Mitsui, "Optimum Design for High-Power and High Efficiency GaAs Hi-Lo IMPATT Diodes," *IEEE Trans. Electron Dev.*, **ED-26**, 210 (1979).
23. W. J. Evans, "Avalanche Diode Oscillators," in W. D. Hershberger, Ed., *Solid State and Quantum Electronics*, Wiley, New York, 1971.
24. T. Misawa, "Saturation Current and Large Signal Operation of a Read Diode," *Solid-State Electron.*, **13**, 1363 (1970).
25. Y. Aono and Y. Okuto, "Effect of Undepleted High Resistivity Region on Microwave Efficiency of GaAs IMPATT Diodes," *Proc. IEEE*, **63**, 724 (1975).
26. B. C. DeLoach, Jr., "Thin Skin IMPATTs," *IEEE Trans. Microwave Theory Tech.*, **MTT-18**, 72 (1970).
27. T. Misawa, "High Frequency Fall-Off of IMPATT Diode Efficiency," *Solid-State Electron.*, **15**, 457 (1972).

28. J. J. Berenz, J. Kinoshita, T. L. Hierl, and C. A. Lee, "Orientation Dependence of *n*-type GaAs Intrinsic Avalanche Response Time," *Electron. Lett.*, **15**, 150 (1979).
29. L. H. Holway, Jr. and M. G. Adlerstein, "Approximate Formulas for the Thermal Resistance of IMPATT Diodes Compared with Computer Calculations," *IEEE Trans. Electron Dev.*, **ED**-24, 156 (1977).
30. M. F. Hines, "Noise Theory for Read Type Avalanche Diode," *IEEE Trans. Electron Dev.*, **ED-13**, 158 (1966).
31. H. K. Gummel and J. L. Blue, "A Small-Signal Theory of Avalanche Noise on IMPATT Diodes," *IEEE Trans. Electron Dev.*, **ED-14**, 569 (1967).
32. J. L. Blue, "Preliminary Theoretical Results on Low Noise GaAs IMPATT Diodes," *IEEE Device Res. Conf.*, Seattle, Wash., June 1970.
33. W. Harth, W. Bogner, L. Gaul, and M. Claassen, "A Comparative Study on the Noise Measure of Millimeter-Wave GaAs IMPATT Diodes," *Solid-State Electron.*, **37**, 427 (1994).
34. J. C. Irvin, D. J. Coleman, W. A. Johnson, I. Tatsuguchi, D. R. Decker, and C. N. Dunn, "Fabrication and Noise Performance of High-Power GaAs IMPATTs," *Proc. IEEE*, **59**, 1212 (1971).
35. L. S. Bowman and C. A. Burrus, Jr., "Pulse-Driven Silicon *p-n* Junction Avalanche Oscillators for the 0.9 to 20 mm Band," *IEEE Trans. Electron Dev.*, **ED-14**, 411 (1967).
36. M. Ino, T. Ishibashi, and M. Ohmori, "Submillimeter Wave Si p^+pn^+ IMPATT Diodes," *Jpn. J. Appl. Phys.*, **16**, Suppl. **16-1**, 89 (1977).
37. J. Pribetich, M. Chive, E. Constant, and A. Farrayre, "Design and Performance of Maximum-Efficiency Single and Double-Drift-Region GaAs IMPATT Diodes in the 3–18 GHz Frequency Range," *J. Appl. Phys.*, **49**, 5584 (1978).
38. T. E. Seidel, R. E. Davis, and D. E. Iglesias, "Double-Drift-Region Ion-Implanted Millimeter-Wave IMPATT Diodes," *Proc. IEEE*, **59**, 1222 (1971).
39. H. Eisele and R. Kamoua, "Submillimeter-Wave InP Gunn Devices," *IEEE Trans. Microwave Theory Tech.*, **52**, 2371 (2004).
40. M. Arai, S. Ono, and C. Kimura, "IMPATT Oscillation in SiC p^+-n^--n^+ Diodes with a Guard Ring Formed by Vanadium Ion Implantation," *Electron. Lett.*, **40**, 1026 (2004).
41. A. K. Panda, D. Pavlidis, and E. Alekseev, "DC and High-Frequency Characteristics of GaN-Based IMPATTs," *IEEE Trans. Electron Dev.*, **ED-48**, 820 (2001).
42. J. K. Mishra, G. N. Dash, S. R. Pattanaik, and I. P. Mishra, "Computer Simulation Study on the Noise and Millimeter Wave Properties of InP/GaInAs Heterojunction Double Avalanche Region IMPATT Diode," *Solid-State Electron.*, **48**, 401 (2004).
43. D. J. Coleman, Jr. and S. M. Sze, "The BARITT Diode—A New Low Noise Microwave Oscillator," IEEE Device Res. Conf., Ann Arbor, Mich., June 28, 1971; "A Low-Noise Metal-Semiconductor-Metal (MSM) Microwave Oscillator," *Bell Syst. Tech. J.*, **50**, 1695 (1971).
44. H. W. Ruegg, "A Proposed Punch-Through Microwave Negative Resistance Diode," *IEEE Trans. Electron Dev.*, **ED-15**, 577 (1968).
45. G. T. Wright, "Punch-Through Transit-Time Oscillator," *Electron. Lett.*, **4**, 543 (1968).
46. S. M. Sze, D. J. Coleman, and A. Loya, "Current Transport in Metal-Semiconductor-Metal (MSM) Structures," *Solid-State Electron.*, **14**, 1209 (1971).

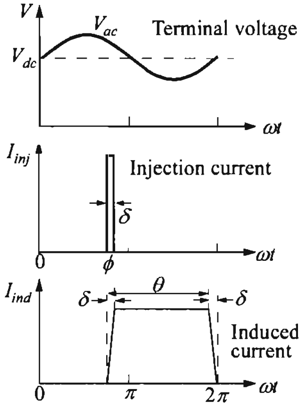
47. J. L. Chu, G. Persky, and S. M. Sze, "Thermionic Injection and Space-Charge-Limited Current in Reach-Through p^+np^+ Structures," *J. Appl. Phys.*, **43**, 3510 (1972).
48. J. L. Chu and S. M. Sze, "Microwave Oscillation in *pnp* Reach-Through BARITT Diodes," *Solid-State Electron.*, **16**, 85 (1973).
49. H. Nguyen-Ba and G. I. Haddad, "Effects of Doping Profile on the Performance of BARITT Devices," *IEEE Trans. Electron Dev.*, **ED-24**, 1154 (1977).
50. S. P. Kwok and G. I. Haddad, "Power Limitation in BARITT Devices," *Solid-State Electron.*, **19**, 795 (1976).
51. O. Eknoyan, S. M. Sze, and E. S. Yang, "Microwave BARITT Diode with Retarding Field—An Investigation," *Solid-State Electron.*, **20**, 285 (1977).
52. J. Nishizawa, "The GaAs TUNNETT Diodes," in K. J. Button, Ed., *Infrared and Millimeter Waves*, Vol. **5**, p. 215, Academic Press, New York, 1982.
53. J. Nishizawa, P. Plotka, H. Makabe, and T. Kurabayashi, "GaAs TUNNETT Diodes Oscillating at 430–655 GHz in CW Fundamental Mode," *IEEE Microwave Wireless Comp. Lett.*, **15**, 597 (2005).

PROBLEMS

1. The idealized voltage and current waveforms for a transit-time diode are shown. The δ is the injected pulse width, ϕ is the phase delay and is at the center of the pulse, and θ is the drift-region angle. The dc current is

$$I_{dc} = \frac{1}{2\pi}\int_0^{2\pi} I_{ind}\,d(\omega t)$$

where I_{ind} is the induced current. Find (a) I_{max} in terms of I_{dc} and θ. (b) The efficiency ($\eta \equiv P_{ac}/P_{dc}$). (c) If $V_{ac}/V_{dc} = 0.5$, $\phi = 3\pi/2$, $\delta = 0$, and $\theta = \pi/2$, find the efficiency.



2. Find the dc reverse bias voltage for a silicon p^+-i-n^+-i-n^+ IMPATT diode operated at 1 A. The thickness of the first i-region is 1.5 μm, the second i-region is 4.5 μm, and the doping in the n^+ δ-region is 10^{18} cm^{-3} with a width of 14 nm. The device area is 5×10^{-4} cm^2. Neglect the temperature effect.

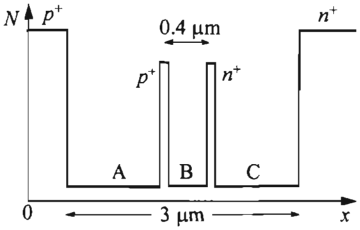
3. For the IMPATT diode in Prob. 2:

(a) when the device is under avalanche breakdown condition, is the field in the drift region high enough to maintain velocity saturation of electrons, and

(b) estimate the maximum cw power output of this device, assuming the diode capacitance is 0.05 pF and Pf^2 = constant.

4. A Si IMPATT diode is operated at 94 GHz with a dc bias of 20 V and an average biasing current of 200 mA. If the power conversion efficiency is 25%, the thermal resistance is 30°C/W, and the breakdown voltage increases with temperature at a rate of 40 mV/°C, find the breakdown voltage of the diode at room temperature (assuming the space-charge resistance is zero).

5. Consider a GaAs double-drift lo-hi-lo IMPATT diode with an avalanche region width of 0.4 μm and a total depletion region width of 3 μm. The n^+ or p^+ clump has a charge of 1.5×10^{12} cm^{-2}. The dopings in regions A, B, and C are assumed to be very low. Find the breakdown voltage of the device.



6. A Si IMPATT diode is operated at 140 GHz with a dc bias of 15 V and an average biasing current 150 mA. (a) If the power conversion efficiency is 25% and the thermal resistance of the diode is 40°C/W, find the junction temperature rise above the room temperature. (b) If the breakdown voltage increases with temperature at a rate of 40 mV/°C, find the breakdown voltage of the diode at room temperature, assuming that the space-charge resistance is 10 Ω, independent of temperature.

7. A transit-time diode has an injection phase delay of $3\pi/2$, a transit angle of $\pi/4$, and a ratio of $V_{ac}/V_{dc} = 0.5$. Find the dc-to-ac power conversion efficiency.

8. A symmetric PtSi-Si-PtSi (MSM) structure has $N_D = 4\times10^{14}$ cm^{-3}, $W = 12$ μm, and an area of 5×10^{-4} cm^2. Find (a) its zero-bias capacitance and (b) the capacitance at reach-through voltage.

9. Based on large-signal operation of a BARITT diode (Fig. 24, p. 504), estimate the microwave power generation efficiency under the condition that the peak ac voltage V_{ac} is equal to $0.4V_{RT}$, and the dc voltage is equal to V_{RT}.

10. Derive Eq. 71.

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

10

Transferred-Electron and Real-Space-Transfer Devices

10.1 INTRODUCTION
10.2 TRANSFERRED-ELECTRON DEVICE
10.3 REAL-SPACE-TRANSFER DEVICES

10.1 INTRODUCTION

In this chapter we introduce two different mechanisms that lead to negative differential resistance—the *transferred-electron effect* and *real-space transfer*. The commonality of these is that under high electric fields, carriers are transferred to a different space in which the carriers have lower mobility and drift velocity. As a result, a reduced current is realized under higher bias, the definition of negative differential resistance. The difference among the two mechanisms is that they occur in completely different kinds of space: transferred-electron effect in k-space of the energy-momentum (E-k) relationship and real-space transfer in different semiconductor materials across a heterointerface. The former is thus a bulk material property while the second is based on heterojunctions of two materials. The transferred-electron effect leads to the transferred-electron device (TED) which is a two-terminal diode. The real-space transfer can result in a two-terminal diode or a three-terminal transistor. The details of these mechanisms and devices will be discussed in their respective sections.

10.2 TRANSFERRED-ELECTRON DEVICE

The *transferred-electron device* (TED), also called Gunn diode, one of the most-important microwave devices, has been used extensively as local oscillators and power amplifiers, covering the microwave frequency range from 1 to 300 GHz. The TEDs have matured to become important solid-state microwave sources used in radars, intrusion alarms, and microwave test instruments.

In 1963, Gunn discovered that repetitive microwave current-pulse output was generated when a dc electric field that exceeded a critical threshold value was applied across a randomly oriented, short, *n*-type sample of GaAs or InP.[1,2] The frequency of oscillation was approximately equal to the reciprocal of the carrier transit time across the length of the sample. Later, Kroemer[3] pointed out that all the observed properties of the microwave oscillation were consistent with a theory of negative differential mobility, independently proposed earlier by Ridley and Watkins[4,5] and by Hilsum.[6,7] The mechanism responsible for the negative differential mobility is a field-induced transfer of conduction-band electrons from a low-energy, high-mobility valley to higher-energy, low-mobility satellite valleys. The GaAs pressure experiments of Hutson et al.[8] and the $GaAs_{1-x}P_x$ alloy experiments of Allen et al.,[9] which demonstrated that the threshold electric field decreases with decreasing energy separation between the valley minima, were convincing evidence that the transferred-electron effect was indeed responsible for the Gunn oscillation.

The transferred-electron effect has also been referred to as the Ridley-Watkins-Hilsum effect or the Gunn effect. Comprehensive reviews on TEDs have been given by Refs. 10–14.

10.2.1 Transferred-Electron Effect

The transferred-electron effect is the transfer of conduction electrons from a high-mobility energy valley to low-mobility, higher-energy satellite valleys. To understand how this effect leads to negative differential resistance (NDR), consider the energy-momentum diagrams for GaAs and InP (Fig. 1), the two most-important semiconductors for TEDs.[15,16] As seen, the band structures of GaAs and InP are very similar. The conduction band consists of a number of subbands. The bottom of the conduction band is located at $\boldsymbol{k} = 0$ (Γ point). The first higher subband is located along the $\langle 111 \rangle$-axis (L), and the next higher subband appears along the $\langle 100 \rangle$-axis (X). Therefore, the ordering of the subbands in both semiconductors is Γ-L-X. Until Aspnes made his synchrotron radiation, Schottky-barrier, electroreflectance measurements in 1976,[15] the first subband in GaAs was generally taken as X with a separation of about 0.36 eV at room temperature. These measurements established the correct ordering of the subbands for GaAs as Γ-L-X, which is identical to that for InP (Fig. 1).

We shall now derive an approximate velocity-field characteristic based on the single-temperature model, that is, electrons in the lower valley (Γ) and the upper valleys (L) are assigned the same electron temperature T_e.[6,17] The energy separation

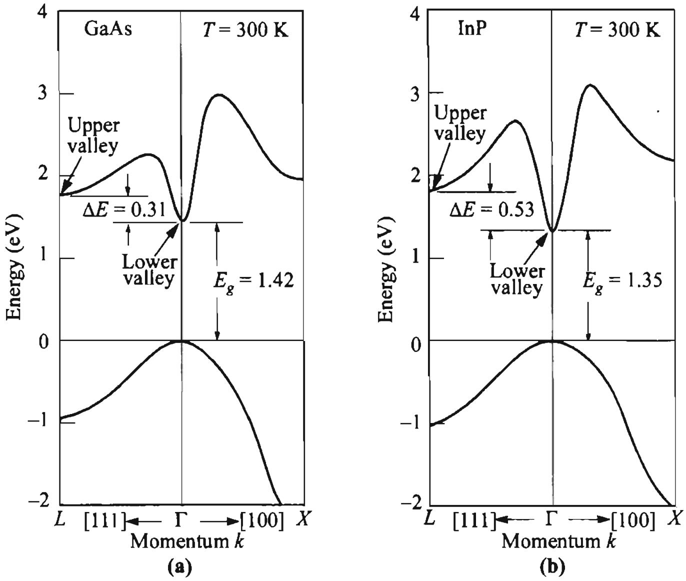


Fig. 1 Energy-band structures of (a) GaAs and (b) InP. The lower conduction valley is at $k = 0$ (Γ); the high valley is along the ⟨111⟩-axis (*L*). (After Refs. 15 and 16.)

between the two valleys ΔE is about 0.31 eV for GaAs and 0.53 eV for InP. The lower-valley effective mass is denoted by m_1^*, and the mobility by μ_1. The same upper-valley quantities are denoted by m_2^* and μ_2, respectively. Also, the densities of electrons in the lower and upper valleys are n_1 and n_2, respectively, with the total concentration $n = n_1 + n_2$. The steady-state current density of the semiconductor may be written as

$$J = q(\mu_1 n_1 + \mu_2 n_2)\mathscr{E} = qnv, \tag{1}$$

where the average drift velocity v is

$$v = \left(\frac{\mu_1 n_1 + \mu_2 n_2}{n_1 + n_2}\right)\mathscr{E} \approx \frac{\mu_1 \mathscr{E}}{1 + (n_2/n_1)} \tag{2}$$

since $\mu_1 \gg \mu_2$. The population ratio between the upper and lower valleys with an energy difference of ΔE is

$$\frac{n_2}{n_1} = R\ \exp\left(-\frac{\Delta E}{kT_e}\right) \tag{3}$$

where R is the density-of-state ratio given by

$$R = \frac{\text{available states in all upper valleys}}{\text{available states in lower valley}} = \frac{M_2}{M_1}\left(\frac{m_2^*}{m_1^*}\right)^{3/2}. \tag{4}$$

The M_l and M_2 are the numbers of equivalent lower and upper valleys, respectively. For GaAs, $M_l = 1$ and there are eight upper valleys in the L-direction, but they happen to be near the edge of the first Brillouin zone, and therefore $M_2 = 4$. Using the effective masses $m_l^* = 0.067m_0$ and $m_2^* = 0.55m_0$, R is found to be 94 for GaAs.

The electron temperature T_e is higher than the lattice temperature T, since the electric field accelerates the electrons and increases their kinetic energy. The electron temperature is determined through the concept of energy relaxation time τ_e:

$$qv\mathscr{E} = \frac{3k(T_e - T)}{2\tau_e} \tag{5}$$

where τ_e is assumed to be of the order of 10^{-12} s. Substituting v from Eq. 2 and n_2/n_l from Eq. 3, we obtain from Eq. 5,

$$T_e = T + \frac{2q\tau_e\mu_l}{3k}\mathscr{E}^2\left[1 + R\exp\left(-\frac{\Delta E}{kT_e}\right)\right]^{-1}. \tag{6}$$

We can compute T_e as a function of the electric field for a given T, and from Eqs. 2 and 3 the velocity-field characteristic can be written as

$$v = \mu_l\mathscr{E}\left[1 + R\exp\left(-\frac{\Delta E}{kT_e}\right)\right]^{-1}. \tag{7}$$

The general v-$\mathscr{E}$ curves as obtained from Eqs. 6 and 7 are shown in Fig. 2 for GaAs with three lattice temperatures. Also shown is the upper-valley population fraction P ($= n_2/n$) as a function of the field.

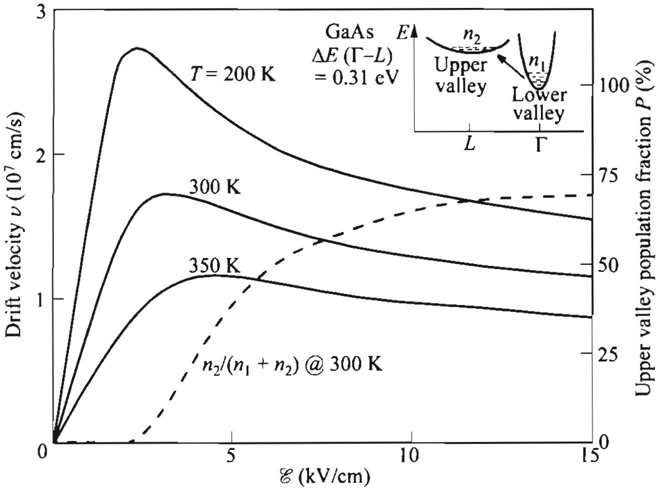


Fig. 2 Calculated velocity-field characteristics for GaAs at three lattice temperatures based on a two-valley model having a single electron temperature.

From the v-$\mathscr{E}$ curves in Fig. 2, the I-V characteristics of the device duplicate exactly the same shapes. As seen, a region of negative differential resistance exists. However, what is unique in a TED is the origin of this NDR, which is the field dependence of drift velocity, as opposed to other mechanisms in a tunnel diode or a real-space-transfer diode. This field-dependent velocity leads to an interesting internal instability that forms the charge domains observed by Gunn as current pulses. Domain formation will be discussed in the next section. Here it is more appropriate to introduce the concept of differential mobility μ_d, defined as

$$\mu_d \equiv \frac{dv}{d\mathscr{E}}. \tag{8}$$

This is different from the conventional low-field mobility ($\mu \equiv v/\mathscr{E}$) that we use in field-effect transistors. So by definition, the low-field mobility is independent of the field, which is not necessarily the case for differential mobility.

In practical operating TEDs, the higher valley has lower mobility and the carriers residing there are driven into velocity saturation under a large field. The average velocity from Eq. 2 can then be modified to

$$\begin{aligned} v &= \frac{n_1\mu_1\mathscr{E} + n_2 v_s}{n_1 + n_2} \\ &= \mu_1\mathscr{E} - P(\mu_1\mathscr{E} - v_s). \end{aligned} \tag{9}$$

The differential mobility is given by

$$\mu_d = \frac{dv}{d\mathscr{E}} = \mu_1(1-P) + (v_s - \mu_1\mathscr{E})\frac{dP}{d\mathscr{E}}. \tag{10}$$

With some mathematical manipulation it can be shown that μ_d is negative under the operation condition of

$$\frac{dP}{d\mathscr{E}} > \frac{1-P}{\mathscr{E} - (v_s/\mu_1)}. \tag{11}$$

The simple models discussed above show the following points: (a) there is a well-defined threshold field ($\mathscr{E}_T$) for the onset of NDR (or negative differential mobility); (b) the threshold field increases with the lattice temperature; and (c) the negative mobility can be destroyed by having the lattice temperature too high or the energy difference ΔE too small. Therefore, certain requirements must be met for the electron-transfer mechanism to give rise to bulk NDR: (1) The lattice temperature must be low enough that, in the absence of a bias electric field, most electrons are in the lower conduction-band minimum, or $kT < \Delta E$. (2) In the lower conduction-band minimum, the electrons must have high mobility, small effective mass, and low density of states; whereas in the upper satellite valleys, the electrons must have low mobility, large effective mass, and high density of states. (3) The energy separation between the two valleys must be smaller than the semiconductor bandgap so that avalanche breakdown does not set in before electrons are transferred into the upper valleys.

Of the semiconductors satisfying these conditions, n-type GaAs and n-type InP are the most widely studied and used. The transferred-electron effect, however, has been observed in many other semiconductors, including Ge, some binary, ternary, and quaternary compounds (see Table 1).[12,18,19] The transferred-electron effect is observed in InAs and InSb under hydrostatic pressures that are applied to decrease the energy difference ΔE, which under normal pressures is greater than the energy gap. Of particular interest are the GaInSb ternary III-V compounds for potential low-power, high-speed applications because of their low threshold fields and high velocities. For semiconductors with large valley energy separations (e.g., $Al_{0.25}In_{0.75}As$ with $\Delta E = 1.12$ eV and $Ga_{0.6}In_{0.4}As$ with $\Delta E = 0.72$ eV) the negative resistance may become dominated by the central Γ-valley:[20] Monte Carlo calculations have shown that for these semiconductors, the presence of the upper valleys is not required for a negative-resistance effect but that polar optical scattering acting in a nonparabolic central valley alone gives rise to a peak velocity and negative-resistance effect.

The measured room-temperature velocity-field characteristics for GaAs and InP are shown in Fig. 3. The analysis based on high-field carrier transport studies are in good agreement with the experimental results.[22,23] The threshold field $\mathscr{E}_T$ defining the onset of NDR is approximately 3.2 kV/cm for GaAs and 10.5 kV/cm for InP. The peak velocity v_p is about 2.2×10^7 cm/s for high-purity GaAs and 2.5×10^7 cm/s for high-purity InP. The maximum negative differential mobility is about -2400 cm²/V-s for GaAs and -2000 cm²/V-s for InP. The measured relative threshold field $\mathscr{E}_T(T)/\mathscr{E}_T(300\text{ K})$ and the relative peak velocity $v_p(T)/v_p(300\text{ K})$ in GaAs as a function of lattice temperature are shown in Fig. 4. The simple model (Fig. 2) is in qualitative agreement with the experimental results.

Table 1 Semiconductor Materials Related to Transferred-Electron Effect at 300 K

		Valley Separation			
Semiconductor	E_g (eV)	Between	ΔE (eV)	$\mathscr{E}_T$ (kV/cm)	v_p (10^7 cm/s)
GaAs	1.42	Γ-L	0.31	3.2	2.2
InP	1.35	Γ-L	0.53	10.5	2.5
Ge[a]	0.74	L-Γ	0.18	2.3	1.4
CdTe	1.50	Γ-L	0.51	11.0	1.5
InAs[b]	0.36	Γ-L	1.28	1.6	3.6
InSb[c]	0.28	Γ-L	0.41	0.6	5.0
ZnSe	2.60	Γ-L	—	38.0	1.5
$Ga_{0.5}In_{0.5}Sb$	0.36	Γ-L	0.36	0.6	2.5
$Ga_{0.3}In_{0.7}Sb$	0.24	Γ-L	—	0.6	2.9
$InAs_{0.2}P_{0.8}$	1.10	Γ-L	0.95	5.7	2.7
$Ga_{0.13}In_{0.87}As_{0.37}P_{0.63}$	1.05	—	—	5.5–8.6	1.2

[a] At 77 K. (100)- or (110)-oriented.
[b] Under 14-kbar pressure.
[c] At 77 K under 8-kbar pressure.

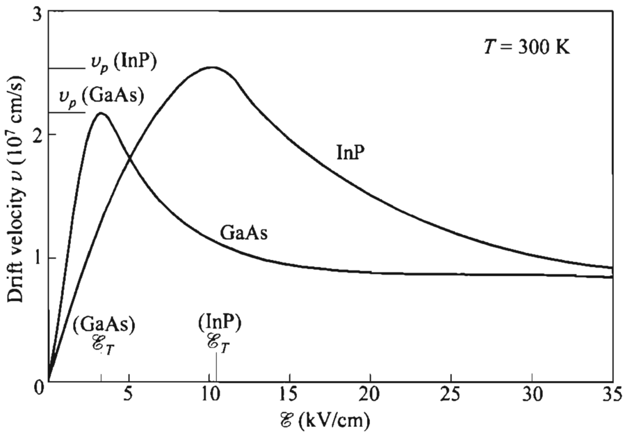


Fig. 3 Measured velocity-field characteristics for GaAs and InP. (After Refs. 16 and 21.)

10.2.2 Domain Formation

Unlike other various physical causes giving rise to negative differential resistance, a semiconductor exhibiting bulk negative differential mobility is inherently unstable, because a random fluctuation of carrier density at any point in the semiconductor produces a momentary space charge that grows exponentially in time. The concept of domain formation and Gunn oscillation is demonstrated qualitatively in Fig. 5. The instability in a TED starts with a dipole which consists of excess electrons (negative

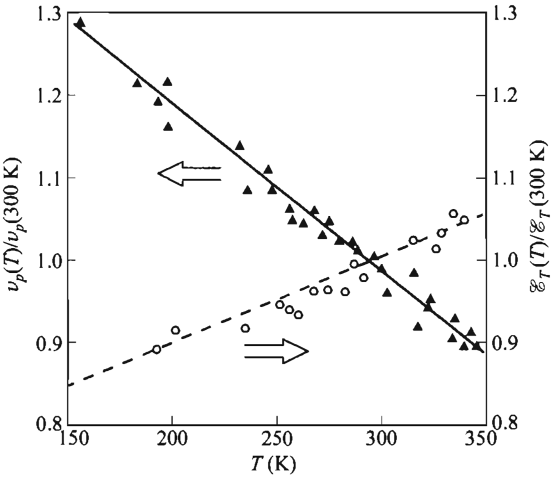


Fig. 4 Measured peak velocity (relative to 300 K) and threshold field (relative to 300 K) in GaAs vs. temperature. (After Ref. 24.)

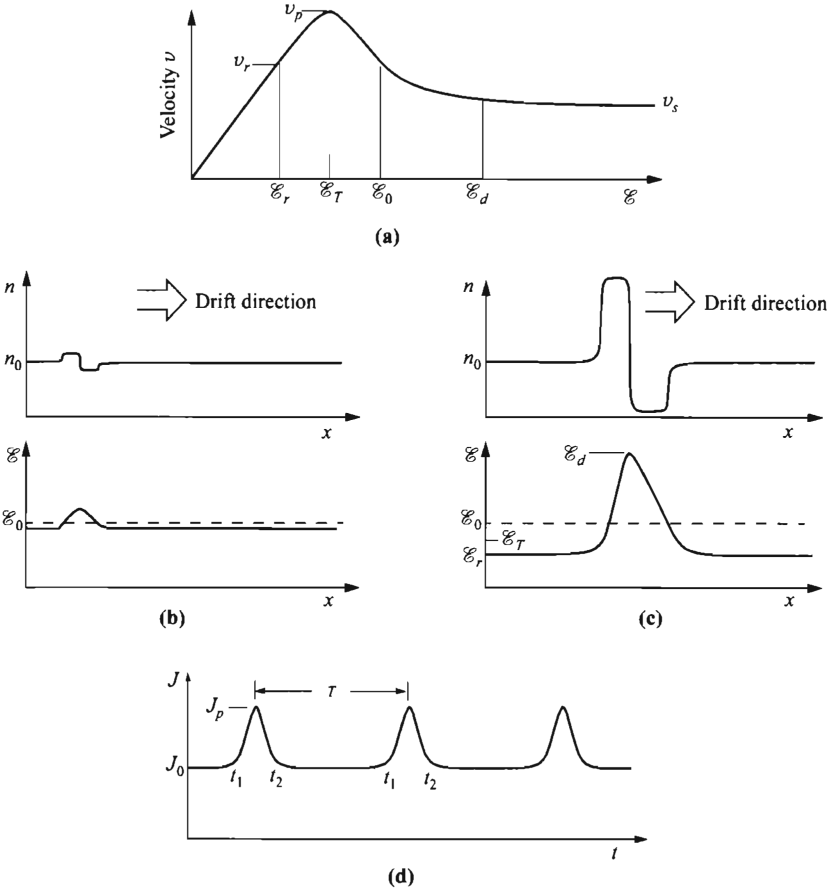

Velocity v
v_p
v_r
v_s
$\mathscr{E}_r$
$\mathscr{E}_T$
$\mathscr{E}_0$
$\mathscr{E}_d$
$\mathscr{E}$
(a)
n
n_0
Drift direction
x
$\mathscr{E}$
$\mathscr{E}_0$
x
(b)
n
n_0
Drift direction
x
$\mathscr{E}$
$\mathscr{E}_d$
$\mathscr{E}_0$
$\mathscr{E}_T$
$\mathscr{E}_r$
x
(c)
J
J_p
J_0
T
t_1
t_2
t_1
t_2
t
(d)

Fig. 5 Demonstration of domain formation. (a) v-$\mathscr{E}$ relationship and some critical points. (b) A small dipole grows to (c) a mature domain. (d) Terminal-current (Gunn) oscillation. Between t_1 and t_2, matured domain is annihilated at the anode and another is formed near the cathode.

charge) and depleted electrons (positive charge) as shown in Fig. 5b. The dipole may arise from many possibilities, such as doping inhomogeneity, material defect, or random noise. This dipole sets up a higher field for the electrons at that location. This higher field, according to Fig. 5a, slows these electrons down relative to the rest outside the dipole. As a result, the region of excess electrons will grow because the trailing electrons behind the dipole are arriving with a higher velocity. By the same token, the region of depleted electrons (positive charge) also grows because electrons ahead of the dipole leave with a higher velocity.

As the dipole grows, the field at that location also increases, but only at the expense of the field everywhere else outside the dipole. The field inside the dipole is always above $\mathscr{E}_0$, and its carrier velocity decreases monotonically with field. The field outside the domain is lower than $\mathscr{E}_0$, and its carrier velocity goes through the peak value and then decreases as the field is lowered. When the field outside the dipole decreases to a certain value, the velocities of electrons inside and outside the dipole are the same (Fig. 5c). At this point the dipole ceases to grow and is said to mature to a *domain*, usually still near the cathode. The domain then transits from near the cathode to the anode.

The terminal-current waveform is shown in Fig. 5d. At t_2, a domain is formed. At t_1, the domain reaches the anode and before another domain is formed, the electric field throughout jumps to $\mathscr{E}_0$. During the formation of a domain $(t_1 - t_2)$, the field outside the dipole passes through the value of $\mathscr{E}_T$ where the peak velocity occurs. This causes a current peak. The current pulse width corresponds to the interval between the annihilation of the domain at the anode and the formation of a new domain. The period T corresponds to the transit time of the domain from cathode to anode.

We now treat the domain formation formally. The one-dimensional continuity equation is given by*

$$\frac{\partial n}{\partial t} + \frac{1}{q}\frac{\partial J}{\partial x} = 0 . \tag{12}$$

If there is a small local fluctuation of the majority carriers from the uniform equilibrium concentration n_0, the locally created space-charge density is $(n - n_0)$. The Poisson equation and the current-density equation are

$$\frac{d\mathscr{E}}{dx} = \frac{q(n-n_0)}{\varepsilon_s} , \tag{13}$$

$$J = \frac{\mathscr{E}}{\rho} - qD\frac{dn}{dx} , \tag{14}$$

where ρ is the resistivity and D the diffusion constant. Differentiating Eq. 14 with respect to x and inserting the Poisson equation yields

$$\frac{1}{q}\frac{dJ}{dx} = \frac{n-n_0}{\rho\varepsilon_s} - D\frac{d^2n}{dx^2} . \tag{15}$$

Substituting this expression into Eq. 12 gives

$$\frac{\partial n}{\partial t} + \frac{n-n_0}{\rho\varepsilon_s} - D\frac{\partial^2 n}{\partial x^2} = 0 . \tag{16}$$

Equation 16 can be solved by separation of variables. For the spatial response, Eq. 16 has the solution

---

* To avoid excessive minus signs, a positive charge will be assigned to electrons and all operations are modified accordingly throughout this chapter.

$$n - n_0 = (n - n_0)\big|_{x=0} \exp\left(\frac{-x}{L_D}\right) \tag{17}$$

where L_D is the Debye length, given by

$$L_D \equiv \sqrt{\frac{kT\varepsilon_s}{q^2 n_0}} \tag{18}$$

which determines the distance over which a small unbalanced charge decays.

For the temporal response, Eq. 16 has the solution:

$$n - n_0 = (n - n_0)\big|_{t=0} \exp\left(\frac{-t}{\tau_R}\right), \tag{19}$$

where τ_R is the dielectric relaxation time given by

$$\tau_R \equiv \rho\varepsilon_s = \frac{\varepsilon_s}{q\mu_d n} \approx \frac{\varepsilon_s}{q\mu_d n_0} \tag{20}$$

which represents the time constant for the decay of the space charge to neutrality if the differential mobility μ_d is positive. However, if the semiconductor exhibits a negative differential mobility, any charge imbalance will grow with a time constant equal to $|\tau_R|$, instead of decay.

Formation of a strong space-charge instability is dependent on the condition that enough charge is available in the semiconductor and the device is long enough that the necessary amount of space charge can be built up within the transit time of the electrons. These requirements set up a criterion for the various modes of operation. In Eq. 19, we have shown that for a device with negative differential mobility, the space charge grows exponentially with a time constant of $|\tau_R| = \varepsilon_s / qn_0|\mu_d|$. If this relationship remained valid throughout the entire transit time of the space-charge layer, the maximum growth factor would be $\exp(L/v_d|\tau_R|)$, where v_d is the average drift velocity of the space-charge layer. For large space-charge growth, this growth factor must be greater than unity, making $L/v_d|\tau_R| > 1$, or

$$n_0 L > \frac{\varepsilon_s v_d}{q|\mu_d|}. \tag{21}$$

For n-type GaAs and InP, the right-hand side of Eq. 21 is about 10^{12} cm^{-2}. TEDs with n_0L products smaller than 10^{12} cm^{-2} exhibit a stable field distribution without current oscillation. Hence, an important boundary separating the various modes of operation is the (carrier concentration) × (device length) product, $n_0 L = 10^{12}$ cm^{-2}.

Domain maturity. The dipole layers will become stable in the sense that they propagate with a particular velocity but do not change in form and size with time. We will assume that the electron drift velocity follows the static velocity-field characteristic shown in Fig. 5a. The equations that determine the behavior of the electron system are the Poisson equation, Eq. 13, and the total current-density equation

$$J = qnv(\mathscr{E}) - q\frac{\partial D(\mathscr{E})n}{\partial x} + \varepsilon_s \frac{\partial \mathscr{E}}{\partial t}. \tag{22}$$

This equation is similar to Eq. 14 except for the addition of the third term, which corresponds to the displacement-current component.

The solutions sought represent a high-field domain that propagates without change of shape, with a domain velocity v_d. Outside the domain, the carrier concentration and fields are at constant values given by $n = n_0$ and $\mathscr{E} = \mathscr{E}_r$, respectively. For this type of solution, both $\mathscr{E}$ and n should be functions of the single variable, $x' \equiv x - v_d t$. Note that n is a double-valued function of field. The domain consists of an accumulation layer where $n > n_0$, followed by a depletion layer where $n < n_0$. The carrier concentration n equals n_0 at two field values, that is, $\mathscr{E} = \mathscr{E}_r$, outside the domain and at $\mathscr{E} = \mathscr{E}_d$, the peak domain field.

Assume that the value of the outside field $\mathscr{E}_r$ is known. (Later it will be shown that $\mathscr{E}_r$ is easily determined.) The current outside the domain consists only of conduction current (given later). Noting that

$$\frac{\partial \mathscr{E}}{\partial x} = \frac{\partial \mathscr{E}}{\partial x'} \tag{23}$$

and

$$\frac{\partial \mathscr{E}}{\partial t} = -v_d \frac{\partial \mathscr{E}}{\partial x'} \tag{24}$$

where v_d is the domain velocity which is the average of carrier velocities inside the domain. One may rewrite Eqs. 13 and 22 as

$$\frac{d\mathscr{E}}{dx'} = \frac{q}{\varepsilon_s}(n - n_0) \tag{25}$$

and

$$\frac{d[D(\mathscr{E})n]}{dx'} = n[v(\mathscr{E}) - v_d] - n_0(v_r - v_d) . \tag{26}$$

We can eliminate the variable x' from these equations by dividing Eq. 26 by Eq. 25 to obtain a differential equation for $[D(\mathscr{E})n]$ as a function of the electric field:

$$\frac{q}{\varepsilon_s}\frac{d[D(\mathscr{E})n]}{d\mathscr{E}} = \frac{n[v(\mathscr{E}) - v_d] - n_0(v_r - v_d)}{n - n_0} . \tag{27}$$

In general, Eq. 27 can only be solved by numerical methods.[25–27] However, the problem may be simplified greatly by assuming that the diffusion term is independent of the electric field, $D(\mathscr{E}) = D$. Using this approximation, the solution to Eq. 27 is given by

$$\frac{n}{n_0} - \ln\left(\frac{n}{n_0}\right) - 1 = \frac{\varepsilon_s}{qn_0D}\int_{\mathscr{E}_r}^{\mathscr{E}} \left\{[v(\mathscr{E}') - v_d] - \frac{n_0}{n}(v_r - v_d)\right\} d\mathscr{E}' . \tag{28}$$

(This solution may be verified by differentiation.)

Note that, when $\mathscr{E} = \mathscr{E}_r$ or $\mathscr{E}_d$, one has $n = n_0$ (Fig. 5c) and the left side of Eq. 28 vanishes; therefore, the integral on the right side of the equation must vanish when $\mathscr{E} = \mathscr{E}_d$. However, the integration from $\mathscr{E}$ to $\mathscr{E}_d$ can represent either the integration over the depletion region when $n < n_0$ or the integration over the accumulation region where $n > n_0$. Since the first term in the integral is independent of n, whereas the contribution from the second term is different in the two cases, one must have $v_r = v_d$, so that the integral vanishes for both integration over the depletion region and over the accumulation region. Then for $\mathscr{E} = \mathscr{E}_d$, Eq. 28 reduces to

$$\int_{\mathscr{E}_r}^{\mathscr{E}_d} [v(\mathscr{E}') - v_r] d\mathscr{E}' = 0 . \tag{29}$$

This equation is satisfied by requiring that the two shaded regions in Fig. 6 have equal areas. By using this rule, the *equal-area rule*,[25] the value of the peak domain field $\mathscr{E}_d$ can be determined if the value of the outside field $\mathscr{E}_r$ is known. The dashed curve in Fig. 6 is a plot of $\mathscr{E}_{\text{dom}}$ against v_r as determined by the equal-area rule. As a function of bias (or $\mathscr{E}_0$), it begins at the peak of the velocity-field characteristic at the threshold field $\mathscr{E}_T$. For outside field ($\mathscr{E}_r$) values resulting in low-field velocities $v(\mathscr{E}_r)$ less than the saturation velocity v_s, the equal-area rule can no longer be satisfied and stable domain propagation cannot be supported.

If the field dependence of the diffusion factor in Eq. 27 is included in the equation, one must use numerical techniques to obtain solutions. These solutions show that for a given value of outside field $\mathscr{E}_r$ there is at most one value of domain excess velocity, defined as $(v_d - v_r)$, for which solutions exist. In other words, only one stable dipole domain configuration exists for each value of $\mathscr{E}_r$.

Now consider some characteristics of a high-field domain. When the domain is not in contact with either electrode, the device terminal current is determined by the outside field $\mathscr{E}_r$ and is given as

$$J_0 = q n_0 v(\mathscr{E}_r) . \tag{30}$$

Therefore, for a given carrier concentration n_0, the outside field fixes the value of J. It is convenient to define the excess voltage on the high-field domain, with outside field $\mathscr{E}_r$, by

$$V_{ex} = \int_{-\infty}^{\infty} [\mathscr{E}(x) - \mathscr{E}_r] dx . \tag{31}$$

The computer solutions of Eq. 31 for different values of carrier concentration and outside field are shown in Fig. 7. These curves may be used to determine the outside

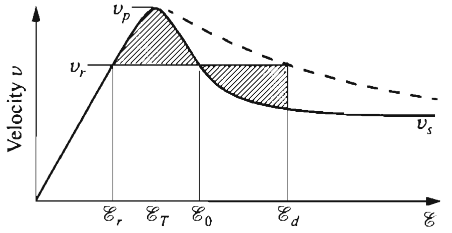


Fig. 6 Velocity vs. field showing the equal-area rule for domain formation. Dashed curve is the locus of v_r vs. $\mathscr{E}_d$ for different domain formation as the bias is varied.

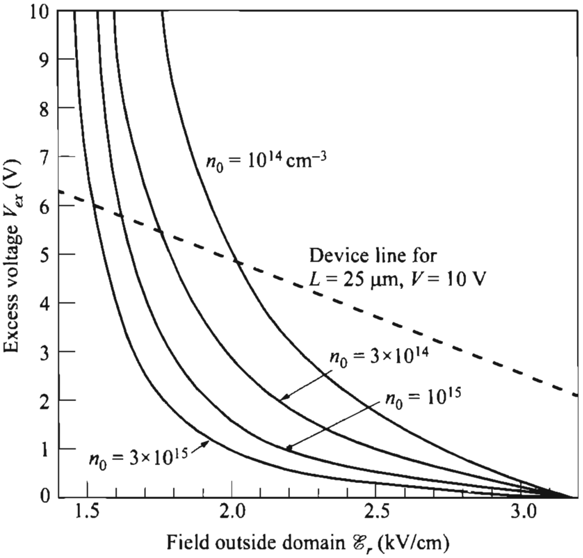


Fig. 7 Excess voltage vs. field for various carrier concentrations. The dashed line is the device line. (After Ref. 27.)

field $\mathscr{E}_r$ in a particular diode of length L, doping concentration n_0, and bias voltage V, by noting that the following relation must hold simultaneously with Eq. 31:

$$V_{ex} = V - L\mathscr{E}_r . \tag{32}$$

The straight line defined by this equation is called the device line and is shown in Fig. 7 as a dashed line for the particular values $L = 25$ μm and $V = 10$ V. If $V/L > \mathscr{E}_T$, the threshold field, the intercept of the device line, and the solutions of Eq. 31 uniquely determine $\mathscr{E}_r$, which in turn specifies the current. The slope of the device line is fixed by L; however, the intercept defining $\mathscr{E}_r$ may be varied by adjusting the bias voltage V.

Figure 8 shows a plot of the domain width versus domain excess voltage.[27] Note that for a given V_{ex}, the domain is narrower for higher doping concentrations. In the limit of zero diffusion, the domain has a triangular shape because when $\mathscr{E}$ in Eq. 28 lies between $\mathscr{E}_r$ and $\mathscr{E}_d$, the right side of the equation approaches infinity as D approaches zero; therefore, the left-hand side must also approach infinity. This requirement implies that $n \rightarrow 0$ in the depletion region and $n \rightarrow \infty$ in the accumulation region. The electric field will vary linearly from $\mathscr{E}_d$ to $\mathscr{E}_r$ and the domain width is

$$d = \frac{\varepsilon_s}{qn_0}(\mathscr{E}_d - \mathscr{E}_r) . \tag{33}$$

The domain excess voltage is then given by

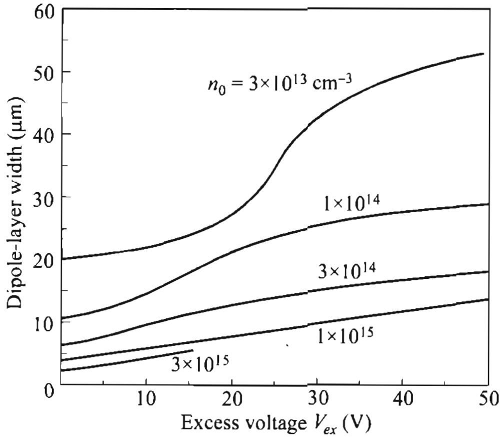


Fig. 8 Domain width vs. domain excess voltage for various doping levels. (After Ref. 27.)

$$V_{ex} = \frac{(\mathscr{E}_d - \mathscr{E}_r)d}{2} = \frac{\varepsilon_s(\mathscr{E}_d - \mathscr{E}_r)^2}{2qn_0}. \tag{34}$$

Experimentally, only triangular domains have been obtained in GaAs and InP TEDs.

When the high-field domain reaches the anode, the current in the external circuit increases and the fields in the diode readjust themselves, nucleating a new domain. Then the frequency of the current oscillations depends on, among other things, the velocity of the domain across the sample, v_d; if v_d increases, the frequency increases, and vice versa. The dependence of v_d on the bias voltage can easily be determined.

When the dipole reaches the anode, the field throughout the sample jumps to a higher value above the threshold field and a new domain is nucleated at the cathode. Figure 9 demonstrates a simulated time-dependent behavior of a domain in a GaAs

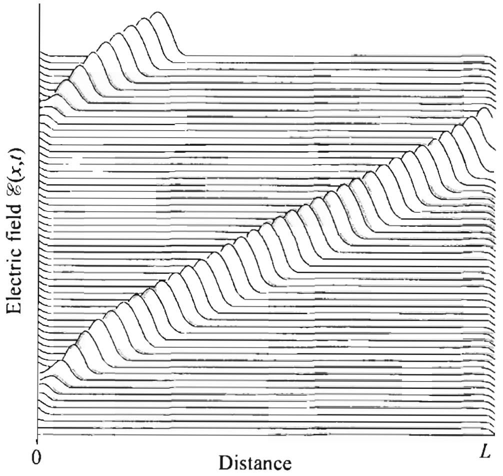


Fig. 9 Numerical simulation of the time-dependent behavior of domain formation and transit. The sample length is 100 μm with doping of 5×10^{14} cm^{-3}. Each successive instant of time is 24 ps. (After Ref. 28.)

device 100-μm long having a doping of 5×10^{14} cm^{-3} ($n_0L = 5\times10^{12}$ cm^{-2}). The time between successive vertical displays of $\mathscr{E}(x,t)$ is $16\tau_R$, where τ_R is the low-field dielectric relaxation time, Eq. 20 ($\tau_R = 1.5$ ps for this device). It is seen here that at any time, only one domain can exist. The terminal current waveform is shown in Fig. 5d. At t_1, the domain reaches the anode. The current pulse then reaches a peak value given by

$$J_p = qn_0v_p. \tag{35}$$

The period of the current pulses is given by the domain transit time (L/v_d). This current oscillation is the first transferred-electron effect observed by Gunn.[1]

10.2.3 Modes of Operation

Since Gunn first observed microwave oscillation in GaAs and InP TEDs in 1963, various modes of operation have been studied. A TED possesses the properties of negative differential resistance based on its I-V characteristics, so the operation can utilize these properties in the same ways as other NDR devices. The additional feature of Gunn current oscillation association with domains can also be used whose frequency is related to the domain transit time. Five major factors affect or determine the modes of operation: (1) doping concentration and doping uniformity in the device, (2) length of the active region, (3) cathode contact property, (4) operating voltage, and (5) type of circuit connected. Different modes of operation will be discussed next.

Ideal Uniform-Field Mode. Under the idealized condition that no internal space charge (domain) has built up and the entire device has a uniform electric field, the current-voltage relationship for a TED can be obtained by scaling the velocity-field characteristics. In this mode of operation, the TED is used as a regular NDR device. Since the operation is not related to the domains, the operating frequency is not restricted to the domain transit time. We consider the simplest voltage waveform of a square wave, as shown in Fig. 10. We shall define two normalization parameters: $\alpha \equiv I_V/I_T$ and $\beta \equiv V_0/V_T$. From the nature of the waveform assumed, the average dc current I_0 is given by

$$I_0 = \frac{(1+\alpha)I_T}{2}. \tag{36}$$

The dc power supplied by the device is

$$P_0 = V_0I_0 = \frac{\beta(1+\alpha)V_TI_T}{2}, \tag{37}$$

and the total RF power available to the load is

$$P_{rf} = \left(\frac{V_M - V_T}{2}\right)\left(\frac{I_T - I_V}{2}\right)\left(\frac{8}{\pi^2}\right) = \frac{(\beta-1)(1-\alpha)V_TI_T}{2}\left(\frac{8}{\pi^2}\right). \tag{38}$$

Therefore, the conversion efficiency from dc to RF is

$$\eta = \frac{(1-\alpha)(\beta-1)}{(1+\alpha)\beta}\left(\frac{8}{\pi^2}\right). \tag{39}$$

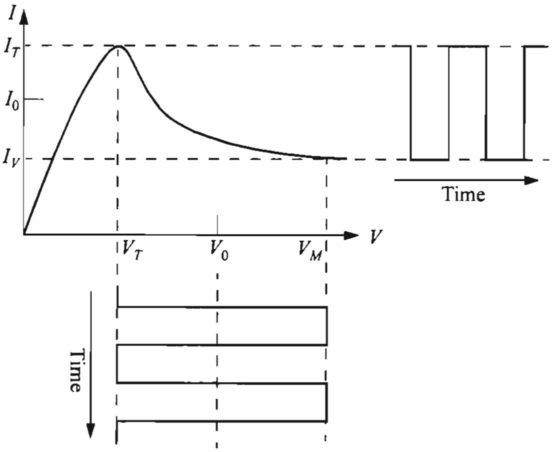


Fig. 10 Idealized square waveforms for the uniform-field mode. V_0 and I_0 are mid-points of ac components. (After Ref. 29.)

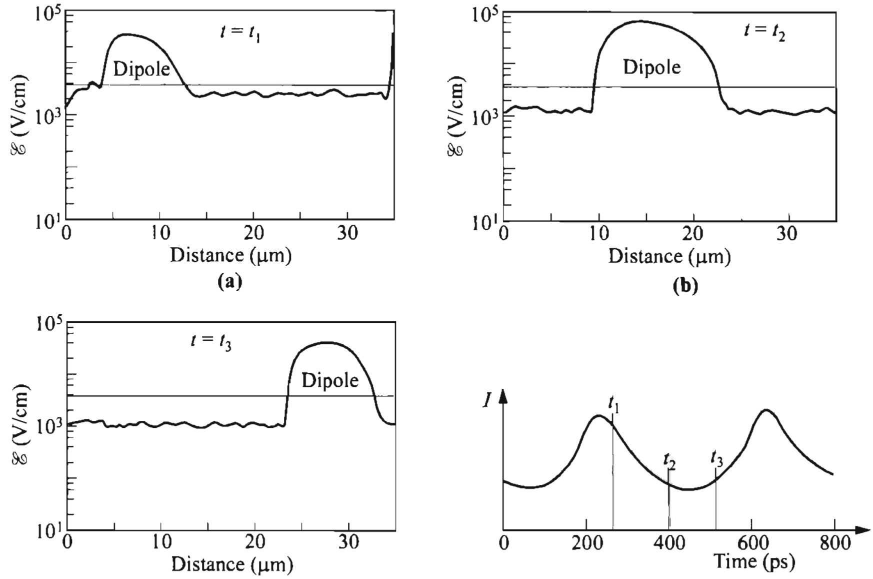
From Eq. 39, we find that maximum efficiency is obtained with a dc bias voltage as high as possible ($\beta \rightarrow \infty$) and with a current peak-to-valley ratio $1/\alpha$ as large as possible. The maximum efficiency yields an ideal value of 30% for GaAs ($1/\alpha = 2.2$) and 45% for InP ($1/\alpha = 3.5$). These efficiencies should be independent of operating frequency as long as the frequency is lower than the reciprocal of the energy relaxation time and the intervalley scattering time.

Experimentally, such high efficiencies have never been obtained and the frequency of operation is generally related to the transit-time frequency, $f = v_d/L$. The reasons are: (1) the bias voltage is limited by avalanche breakdown; (2) a space-charge layer usually forms, giving rise to a nonuniform field, and (3) the ideal current and voltage waveforms are difficult to achieve in a resonant circuit.

Transit-Time Dipole-Layer Mode. When the n_0L product is greater than 10^{12} cm^{-2}, the space-charge perturbations in the material increase exponentially in space and time to form mature dipole layers that propagate to the anode. The dipole is usually formed near the cathode contact, since the largest doping fluctuation and space-charge perturbation exists there. The cyclic formation and subsequent disappearance at the anode of the fully developed dipole layers are what give rise to the experimentally observed Gunn oscillations.

When a TED with overcritical n_0L product is connected in a parallel resonant circuit, such as a high-Q microwave cavity, the transit-time dipole-layer mode can be obtained. In this mode, the high-field domain is nucleated at the cathode and travels the full length of the sample to the anode. Each time a domain is absorbed at the anode, the current in the external circuit increases; therefore, for samples in which the width of the domain is considerably smaller than the length of the sample, the current waveform tends to be spiky rather than the desired sinusoidal form. To obtain a more sinusoidal current waveform, one may either minimize the length of the sample (which increases the frequency in this mode) or increase the width of the domain. Figure 8 shows that the domain width increases with decreasing doping level n_0. In general, more-sinusoidal waveforms may be obtained by decreasing the n_0L product, as long as it exceeds the critical value. Figure 11 shows a sequence of field distributions across a 35-μm sample during one RF cycle, together with the current waveform.[30] The n_0L product is 2.1×10^{12} cm^{-2} and the current waveform is very close to sinusoidal for this device. Theoretical studies show that the efficiency of the transit-time mode is greatest when the n_0L product is one to several times 10^{12} cm^{-2}, so that the domain fills about half the sample length and the current waveform is almost a sine wave. The maximum dc-to-RF conversion efficiency for this mode is 10%. The efficiency can be improved if the current waveform is close to a square wave. This waveform can be produced by adjusting the voltage to be below the threshold at the instant the dipole disappears at the anode. The formation of a new dipole is delayed until the voltage rises above threshold. However, for this delayed domain mode, the tuning procedure is extremely complicated.

Quenched Dipole-Layer Mode. A TED in a resonant circuit can operate at frequencies higher than the transit-time frequency if the high-field dipole layer is quenched before it reaches the anode. In the transit-time dipole-layer mode of operation, most of the voltage across the device is dropped across the high-field dipole layer itself.



(a) (b)

(c) (d)

Fig. 11 Transit-time dipole-layer mode with design for efficiency. Note the large relative domain width and near-sinusoidal current waveform. The GaAs sample has $n_0L = 2.1\times10^{12}$ cm ^{2} and $fL = 0.9\times10^7$ cm/s. (After Ref. 30.)

Therefore, as the resonant circuit reduces the bias voltage, the width of the dipole layer is reduced (Fig. 8). The dipole-layer width continues to decrease as the bias voltage decreases until at some point the accumulation layer and the depletion layer neutralize each other. The bias voltage at which this occurs is designated V_s. Dipole-layer quenching occurs when the bias voltage across the device is reduced below V_s. When the bias voltage swings back above threshold, a new dipole layer is nucleated, and the process repeats. Therefore, the oscillations occur at the frequency of the resonant circuit rather than the transit-time frequency.

Figure 12 shows an example of the quenched dipole-layer mode.[28] The device has identical length and doping as that of Fig. 9. The dipole layer is quenched at a distance of about $L/3$ from the cathode, and the operating frequency is about three times higher than the transit-time dipole-layer mode shown in Fig. 9.

The upper frequency limit for this mode is determined by the speed of quenching, which in turn is determined by two time constants. The first is the positive dielectric relaxation time and the second is an RC time constant; R being the positive resistance in those regions of the diode not occupied by dipoles and C the capacitance of all the dipoles in series. The first condition gives a minimum critical n_0/f ratio of about 10^4 s-cm^{-3} for n-type GaAs and InP.[31,32] The second time constant depends on the number of dipoles and sample length. The efficiency of quenched dipole-layer oscillators can theoretically reach 13%.[33]

In the quenched dipole-layer mode of operation, it has been found both theoretically[30] and experimentally[34] that for samples in which the resonant frequency of the circuit is several times the transit-time frequency (i.e., $fL > 2\times10^7$ cm/s), and the frequency of operation is of the order of the dielectric relaxation frequency (i.e., $n_0/f \approx \varepsilon_s/q|\mu_d|$, as given in Eq. 40), multiple high-field dipole layers usually form. This is due to that one dipole does not have enough time to readjust and absorb the voltage of the other dipoles.

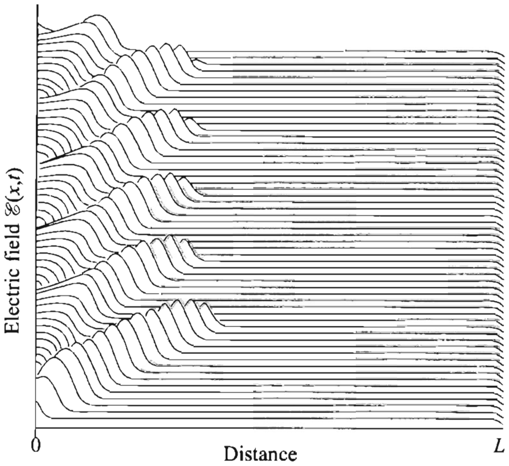


Fig. 12 Numerical simulation of TED in quenched dipole-layer mode. (After Ref. 28.)

Accumulation-Layer Mode. The main difference of the accumulation-layer mode from the dipolar-layer mode is that in a lightly doped or short sample ($n_0L < 10^{12}$ cm^{-2}), it exhibits only accumulation of electrons without a depleted region of positive charge. The consequence is that the field profile becomes a step function around the charge packet, as opposed to the peak shown in Fig. 5c. When a uniform field is applied to such a device, the accumulation-layer dynamics can be understood in a simplified manner, as shown in Fig. 13. At time t_1, an accumulation layer (i.e., excess electrons) is injected from the cathode so that the field distribution splits into two parts, as illustrated at time t_2. The velocities on both sides of the accumulation layer have been altered in the direction shown in Fig. 13a. Since the terminal voltage is assumed to be constant, the area under each electric-field curve of Fig. 13c should be equal. As the accumulation layer propagates toward the anode, this

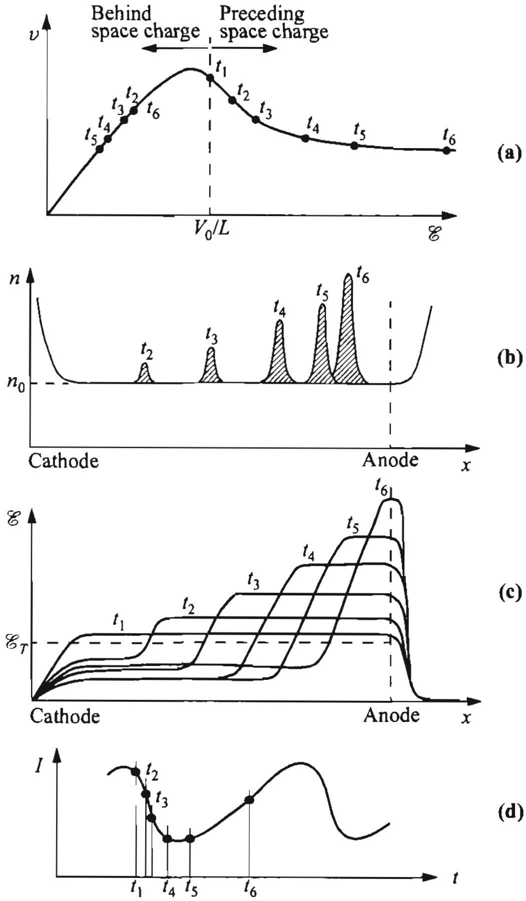
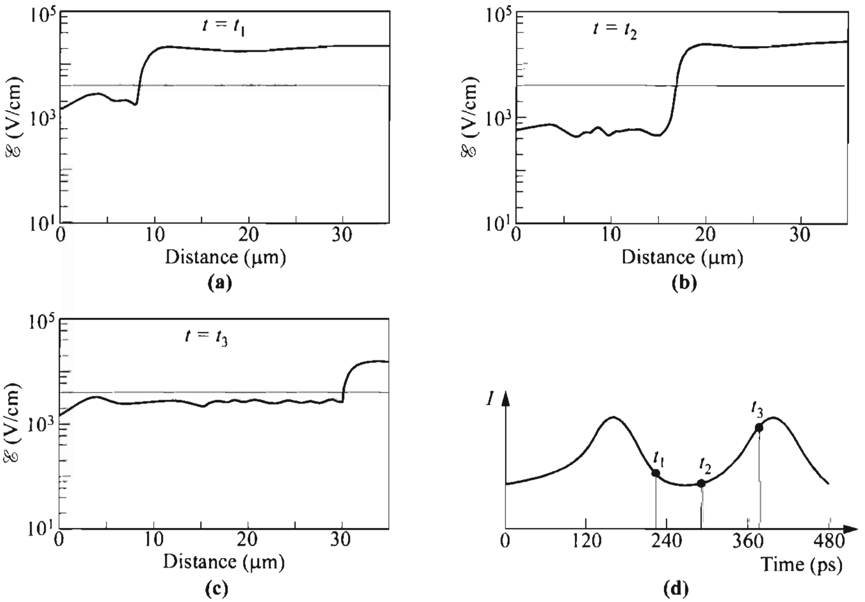


Fig. 13 Accumulation-layer transit mode under time-invariant terminal voltage. (After Ref. 35.)

equality can only be maintained if the velocities on both sides of the accumulation layer fall, as dictated by the velocity-field curve and indicated at times t_3, t_4, and t_5. Eventually, the accumulation layer reaches the anode at time t_6, and disappears there. The field near the cathode rises through the threshold, another accumulation layer is injected, and the process repeats. The smooth current waveform is shown in Fig. 13d. In this particular example, the accumulation charge continues to grow throughout the device length.

A TED with subcritical n_0L product (i.e., $n_0L < 10^{12}$ cm^{-2}) can exhibit negative resistance in a band of frequencies near the electron transit-time frequency and its harmonics. It can be operated as a stable amplifier.[36] When it is connected to a parallel resonant circuit with a load resistor of the order of $10R_0$, where R_0 is the low-field resistance of the TED, it will oscillate in the transit-time accumulation-layer mode. Figure 14 shows the electric field versus distance at three different times during one RF cycle.[30] Also shown is the terminal current waveform. The voltage is always above the threshold value ($V > V_T = \mathscr{E}_T L$). These waveforms are far from ideal; the efficiency for this particular waveform is only 5%. More favorable waveforms with about 10% efficiency can be obtained if the TED is connected to a series resistor and inductor. In this example, the space charge ceases to grow as it drifts to the anode.

Limited-Space-Charge Accumulation (LSA) Mode. In the model of the LSA mode of operation,[31] the electric field across the device rises from below the



(a) (b) (c) (d)

Fig. 14 (a)–(c) Electric field vs. distance at three intervals of time during one RF cycle in accumulation-layer mode. (d) Terminal-current waveform in a resonant circuit for GaAs TED with $n_0 = 2\times10^{14}$ cm^{-3}, $fL = 1.4\times10^7$ cm/s, and $n_0/f = 5\times10^4$ s-cm^{-3}. (After Ref. 30.)

threshold and falls back again so quickly that it is assumed the space-charge distribution associated with high-field dipole layers does not have sufficient time to form. Only the primary accumulation layer forms near the cathode; and the rest of the device remains fairly homogeneous, provided that doping fluctuations are sufficiently small to prevent the formation of dipole layers. Under these conditions, a large portion of the device exhibits a uniform field, which yields efficient power generation at the circuit-controlled frequency. The higher the frequency, the shorter the distance the space-charge layer travels, leaving most of the device biased in the negative-mobility range. The conditions for this mode of operation are derived from two requirements: the space charge should not have enough time to grow to an appreciable size and the accumulation layer must be quenched completely during one RF cycle. Thus, the negative τ_R, Eq. 20, should be larger than the RF period whereas the positive τ_R should be smaller. These requirements lead to the condition[32]

$$\frac{\varepsilon_s}{qn_0\mu_{d+}} \ll \frac{1}{f} < \frac{\varepsilon_s}{qn_0|\mu_{d-}|} \tag{40}$$

where μ_{d+} is the positive differential mobility at low field, and μ_{d-} an average negative differential mobility above the threshold field. For GaAs and InP the two limiting ratios are

$$10^4 < \frac{n_0}{f} < 10^5 \text{ s-cm}^{-3}. \tag{41}$$

It is interesting to note that the quenched multiple dipole-layer mode also occurs in some range of n_0/f ratios if doping fluctuations are present. The LSA device is suited for generating short pulses of high-peak power because devices with overlengths (nontransit-time mode) can be used for which heat extraction becomes difficult. However, the maximum operating frequency of LSA devices is much lower than that of transit-time devices. This lower frequency is caused by the slow energy relaxation of electrons in the lower valley, leading to increased quenching times. Computer simulations indicate that a minimum time of about 20 ps is required to stay below the threshold voltage for cyclic operation in GaAs; the corresponding upper-frequency limit is about 20 GHz.[37,38] For InP a higher upper frequency is expected.

10.2.4 Device Performances

Cathode Contacts. The TEDs require extremely pure and uniform materials with a minimum of deep donor levels and traps, especially if quenching of space charge is involved in the operation. The first TEDs were fabricated from bulk GaAs and InP with alloyed ohmic contacts. Modern TEDs almost always use epitaxial layers on n^+-substrates deposited by advanced epitaxy techniques such as molecular-beam epitaxy. Typical donor concentrations range from 10^{14} to 10^{16} cm^{-3}, and typical device lengths range from a few microns to several hundred microns. The TED chips are mounted in microwave packages. These packages, as well as related heat sinks, are similar to those for IMPATT diodes.

To improve device performance, injection-limited cathode contacts have been used instead of the n^+-ohmic contacts.[39–41] By using an injecting-limited contact, the threshold field for the cathode current can be adjusted to a value approximately equal to the threshold field $\mathscr{E}_T$ for the onset of NDR. Thus a uniform electric field can result. For ohmic contacts, the accumulation or dipole layer grows within some distance from the cathode due to finite heating time of the lower-valley electrons. This *dead zone* may be as large as 1 μm, which imposes a constraint on the minimum device length and hence on the maximum operating frequency. In an injecting-limited contact, hot electrons are injected from the cathode, reducing the length of the dead zone. Because the transit-time effect can be minimized, the device can exhibit a frequency-independent negative conductance shunted by its parallel-plate capacitance. If an inductance and a sufficiently large conductance are connected to this device, it can be expected to oscillate in a uniform-field mode at the resonant frequency. The theoretical efficiency can then be derived as in Section 10.2.3.

Two classes of injecting-limited contacts have been studied; one is a Schottky barrier with low barrier height, the other is a two-zone cathode structure. Figure 15 compares these contacts with an ohmic contact. For an ohmic contact (Fig. 15a) there is always a low-field region near the cathode, and the field is nonuniform across the device length. For a Schottky barrier under reverse bias, a reasonably uniform field can be obtained (Fig. 15b).[42] The reverse current is given by (see Chapter 3)

$$J_R = A^{**}T^2 \exp\left(\frac{-q\phi_B}{kT}\right) \tag{42}$$

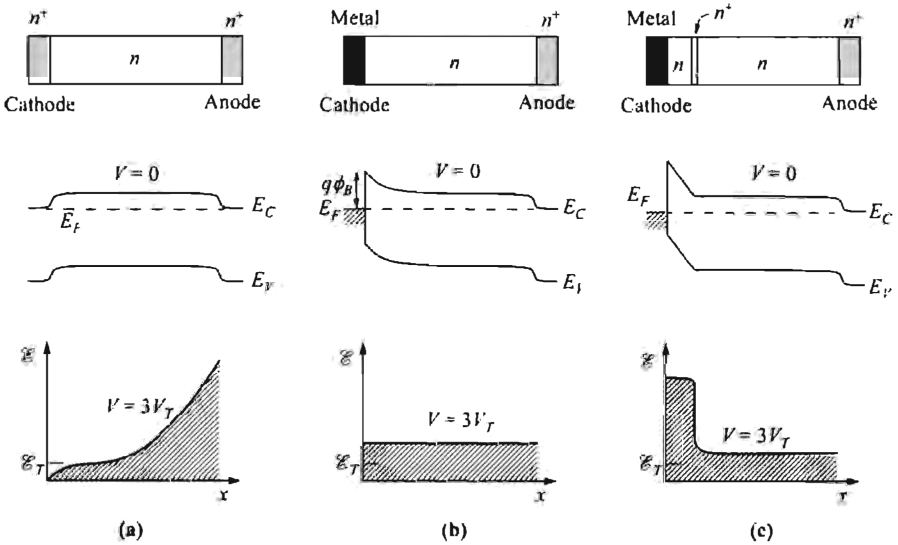


Fig. 15 Three cathode contact schemes: (a) ohmic contact, (b) Schottky-barrier contact, and (c) two-zone Schottky-barrier contact.

where A^{**} is the effective Richardson constant and ϕ_B is the barrier height. For current densities in the range 10^2 to 10^4 A/cm^2, the corresponding barrier height is about 0.15 to 0.3 eV. A Schottky barrier with a low barrier height is not simple to realize in III-V semiconductors and it suffers from a rather limited temperature range because the injected current varies exponentially with temperature (Eq. 42).

The two-zone cathode contact consists of a high-field zone and an n^+-zone (Fig. 15c).[43] This configuration is similar to that of a lo-hi-lo IMPATT diode (see Chapter 9). Electrons are heated in the high-field zone and subsequently injected into the active region having a uniform field. This structure has been successfully used over a wide temperature range.

Power-Frequency Performance and Noise. The transfer of energy from the field to electrons and the scattering of electrons between lower and upper valleys take finite times. These finite times lead to an upper frequency limit corresponding to the scattering and energy relaxation frequencies. Figure 16 shows the time responses of the lower-valley and upper-valley velocities, the upper-valley population fraction, and the average velocity when the electric field is suddenly lowered from 6 kV/cm to

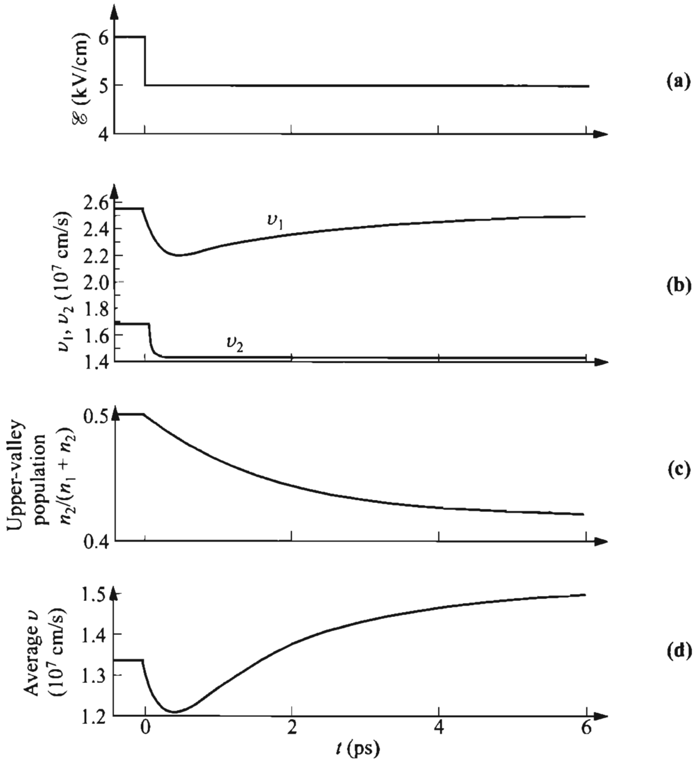


Fig. 16 Response of electrons in upper valley (υ_2, n_2) and lower valley (υ_1, n_1) to an electric field stepped from 6 to 5 kV/cm. (After Ref. 44.)

5 kV/cm. Note that the velocity v_2 of the upper valley follows almost instantaneously with the field. However, the velocity v_1 of the lower valley has a slow response with a time constant of about 2 ps. This response indicates the weak scattering of hot electrons in the lower valley. In addition, the slow decay of n_2 corresponds to the slow scattering from the upper valley to the lower valley. The response of the average velocity v is thus due partly to the recovery of v_1 and partly to the intervalley transfer. Because of the finite response times, the TED has an estimated upper-frequency limit around 500 GHz.

Under transit-time conditions, the frequency of operation is inversely proportional to the device length, that is, $f = v/L$. The power-frequency relationship is given by

$$P_{rf} = \frac{V_{rf}^2}{R_L} = \frac{\mathscr{E}_{rf}^2 L^2}{R_L} = \frac{\mathscr{E}_{rf}^2 v^2}{R_L f^2} \propto \frac{1}{f^2} \tag{43}$$

where V_{rf} and $\mathscr{E}_{rf}$ are the RF voltage and the corresponding field, respectively and R_L is the load impedance. Therefore, the output power is expected to fall as $1/f^2$. The state-of-the-art microwave power versus frequency is shown in Fig. 17 for cw GaAs and InP TEDs. The numbers near the data points indicate conversion efficiencies.

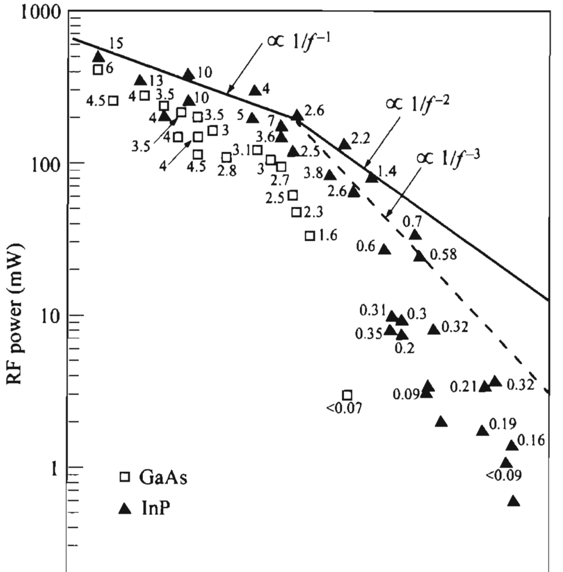


Fig. 17 Output microwave power vs. frequency for cw-operated GaAs and InP TEDs. Numbers next to symbols denote dc-to-RF conversion efficiencies in percentage. (After Ref. 45.)

Note that the power generally varies as $1/f^2$, as given by Eq. 43. The superior performance of InP is apparent, especially for higher frequencies. Generally the cw power is lower than that of an IMPATT diode. On the other hand, the applied voltage for TEDs at a given frequency can be considerably lower than that in an IMPATT diode (by a factor of about 2 to 5) and TEDs have better noise performance.

The TEDs have two types of noise: amplitude deviations (AM noise) and frequency deviations (FM noise), both caused by thermal velocity fluctuations of the electrons. The AM noise is generally small because the amplitude is relatively stable, owing to the strong nonlinearity of the velocity-field characteristic. The FM noise has a mean-frequency deviation given by[46]

$$f_{\mathrm{rms}} = \frac{f_0}{Q_{ex}}\sqrt{\frac{kT_{eq}(f_m)B}{P_0}} \tag{44}$$

where f_0 is the carrier frequency, Q_{ex} the external quality factor, P_0 the power output, and B the measured bandwidth. The modulation-frequency-dependent equivalent noise temperature T_{eq} is given by

$$T_{eq}(f_m) = \frac{qD}{k|\mu_{d-}|} \tag{45}$$

where the average negative differential mobility μ_{d-} depends on the voltage swing. Since the ratio $D/|\mu_{d-}|$ is smaller in InP than in GaAs, the noise is expected to be lower in InP.

Functional Devices. Thus far we have considered the transferred-electron effect and its application to microwave oscillators and amplifiers. TEDs can also be used for high-speed digital and analog operations. We shall consider a TED with nonuniform cross-sectional area or/and nonuniform doping profile, and a three-terminal TED.

The theory of high-field domains in one dimension can be used to analyze nonuniformly shaped oscillators, if one assumes very thin high-field domains and considers phenomena in practically uniform regions in their neighborhood. These assumptions are valid if $n_0L \gg 10^{12}$ cm^{-2} and the variations of the cross-sectional area and doping are gradual. Using the theory presented in the preceding section, it can be shown that there exists a value of domain excess voltage V_{ex} above which the outside electric field $\mathscr{E}_r$ remains constant with time. The value of the average velocity outside the domain corresponding to $\mathscr{E}_r$ is v_r, as shown in Fig. 5a. Such saturated domains move in the oscillator with a constant velocity v_r. The current density associated with a mature domain is given by Eq. 30. For TEDs with nonuniform doping density $N(x)$ and cross-sectional area $A(x)$, however, Eq. 30 can be generalized to

$$I(t) = qN(x)A(x)v(\mathscr{E}_r), \tag{46}$$

where x is measured from the cathode and $x = v_r t$. If a high-field domain is nucleated from the cathode at time $t = 0$, then at time t the domain is at $x(t) = v_r t$ using the foregoing assumptions.

Figure 18 shows the waveform of a bulk-effect oscillator for a sample of nonuniform shape shown.[47] The experimental current waveform is indeed similar to the

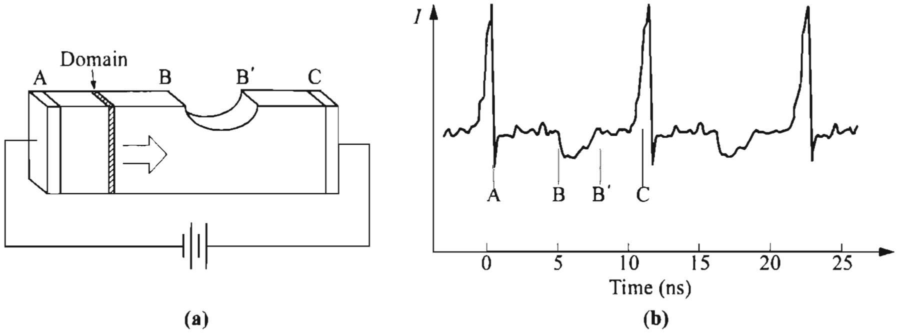


Fig. 18 (a) A TED with nonuniform cross-section and (b) its current waveform. Labels in (b) correspond to the location of domain at the specific time. (After Ref. 47.)

shape of the sample, apart from the known current spikes when the domains reach the anode. The time scale is marked with letters A, B, B′, and C which correspond to the instants of time when the domain is located at these points.

The phenomenon that terminal-current waveform follows that of Eq. 46 can be explained qualitatively as follows. Since the current of the device must be constant at all location, when the domain enters a region of lower doping or smaller cross section, the field of the domain becomes larger in order to maintain similar velocity. A higher domain field (or excess voltage V_{ex}) means the field outside the domain $\mathscr{E}_r$ is lowered. As a consequence, the terminal current is lowered since the field outside the domain determines the current.

Until now, only two-terminal devices have been considered. The current waveform of a TED may be controlled by adding one or more electrodes along the length of the device. Figure 19a shows the structure of such a device with the electrode located at point B. The expected current waveform is shown in Fig. 19b. This waveform can be explained as follows. (The saturated domain theory described previously is used here again.) When the domain leaves the cathode at time $t = 0$, the cathode

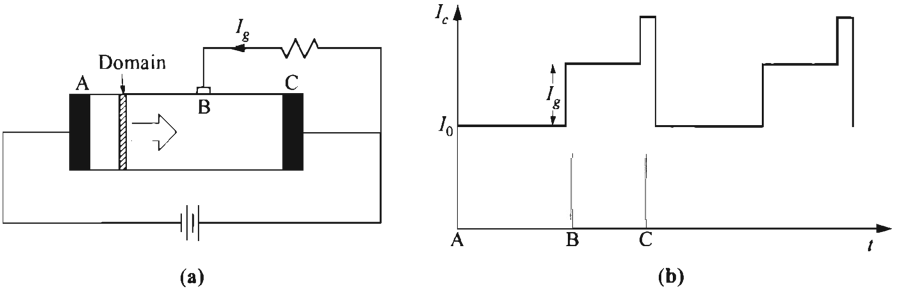


Fig. 19 (a) Circuit of TED controlled-current step generator, and (b) its cathode current waveform. In (b), letters in time scale indicate location of domain. (After Ref. 47.)

current $I_c(t)$ is equal to the current of the saturated domain (Aqn_0v_r), and remains at this value until the domain reaches the electrode at B. At that time the cathode current becomes the sum of the saturated domain current and I_g, the current flowing through the resistor. The current I_g is equal to the voltage sustained by the sample between B and C with a domain present, divided by the resistor value. The cathode current then remains at

$$I_c(t) = Aqn_0v_r + I_g \tag{47}$$

until the domain is absorbed at the anode, at which time the current spikes briefly.

10.3 REAL-SPACE-TRANSFER DEVICES

10.3.1 Real-Space-Transfer Diode

The concept of the *real-space-transfer* (RST) *diode* to obtain negative differential resistance (NDR) was conceived by Gribnikov[48] in 1972 and by Hess et al. in 1979.[49] The first experimental evidence of the negative resistance from a RST diode was shown by Keever et al. in 1981.[50] The requirement of a real-space-transfer diode is a heterostructure whose two materials have different mobilities. In addition, for an *n*-channel device, the material having lower mobility must also have a higher conduction-band edge E_C. A good example is the GaAs/AlGaAs heterostructure.

The real-space-transfer effect[51] is similar to the transferred-electron effect and it is sometimes difficult to separate them experimentally. The transferred-electron effect is due to the properties of a single, homogeneous material. When carriers are excited by a high applied field to a satellite band in the energy-momentum space, the mobility is lower and the current is reduced, resulting in NDR. In real-space transfer, transfer of carriers is between two materials (in real space) rather than two energy bands (in momentum space). In low fields, electrons (in an *n*-channel device) are confined to material (GaAs) with low E_C and higher mobility. The energy-band diagram under high field is shown in Fig. 20. Carriers in the GaAs channel acquire enough energy from the field to overcome the conduction-band discontinuity and flow to the adjacent material (AlGaAs) of lower mobility. This carrier transfer can be considered as thermionic emission, with the electron temperature replacing the lattice temperature. Thus, a higher field results in a smaller current, the definition of NDR. Experimental *I-V* characteristics are shown in Fig. 21. The critical field for this real-space transfer has been shown to be between 2 and 3 kV/cm, while that for the transferred-electron effect is typically 3.5 kV/cm for GaAs. One has to bear in mind that these critical fields are obtained from two different types of channels (heterointerface vs. bulk) and cannot be used alone to separate the effects. Another property of real-space transfer is that there is better control with factors such as conduction-band discontinuity, mobility ratio, and film thicknesses, so that device characteristics can be varied and optimized. The RST effect produces *I-V* characteristics very similar to that of the transferred-electron effect. To achieve an efficient RST diode, a proper choice of heterojunction with optimum band-edge discontinuity, together with a high satellite

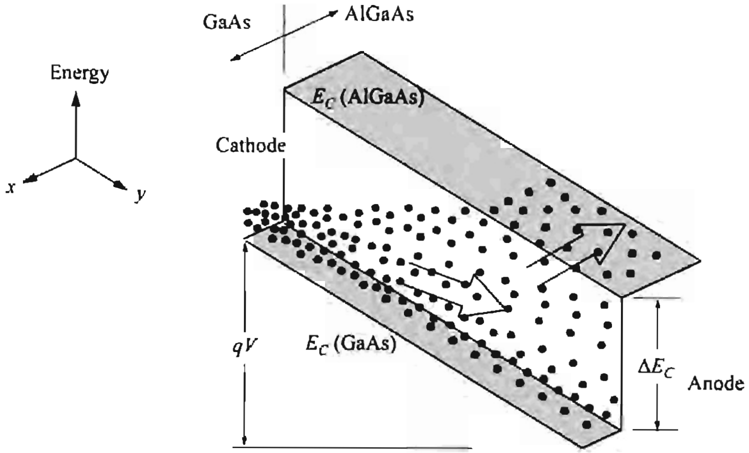


Fig. 20 Energy-band diagram showing the conduction-band edge E_C of the RST diode under bias. Electrons in the main GaAs channel acquire energy from the field to overcome the barrier to spill over to the AlGaAs layer.

valley to avoid the transferred-electron effect (or the absence of a satellite valley), are desirable.

The modeling of the RST diode is complicated, and there are no simple equations derived explicitly for the exact *I-V* characteristics. Qualitatively, the following expressions can be used to get some insight into the origin of the negative resistance. Assume that the total carrier density per unit area is N_s, distributed between the GaAs channel layer of thickness L_1 (n_{s1}) and AlGaAs layer L_2 (n_{s2}):

$$n_{s1} + n_{s2} = N_s . \tag{48}$$

This also implies that carriers can move between the two layers easily, other than having to overcome the barrier ΔE_C when going from the GaAs channel of lower energy to that of higher-energy AlGaAs. The ratio of the carrier densities of the two

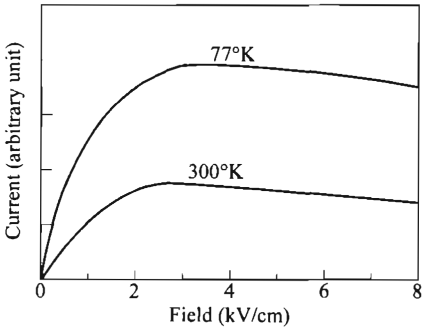


Fig. 21 Current-voltage (or -field) characteristics of a RST diode, based on GaAs/AlGaAs heterostructure. (After Ref. 50.)

layers is related by the energy of the hot carriers, measured by the electron temperature T_e, in relation to the barrier ΔE_C:

$$\frac{n_{s2}}{n_{s1}} = \left(\frac{m_2^*}{m_1^*}\right)^{3/2} \exp\left(\frac{-\Delta E_C}{kT_e}\right). \tag{49}$$

The electron temperature is related to the field by

$$\frac{3k(T_e - T_o)}{2\tau_e} = q\mu_1 \mathscr{E}^2, \tag{50}$$

where τ_e is the energy relaxation time and T_o the lattice (room) temperature.

The fraction of carriers excited to the AlGaAs layer is defined as

$$F(\mathscr{E}) \equiv \frac{n_{s2}}{N_s} \tag{51}$$

and is a function of the applied field. It starts at zero at low field and approaches the ratio of $L_2/(L_1 + L_2)$ at high fields. The total drift current is given by

$$\begin{aligned} J &= qn_{s1}\mu_1\mathscr{E} + Aqn_{s2}\mu_2\mathscr{E} \\ &= q\mathscr{E}N_s[\mu_1 - (\mu_1 - \mu_2)F]. \end{aligned} \tag{52}$$

The differential resistance is given by

$$\frac{dJ}{d\mathscr{E}} = qN_s\left[\mu_1 - (\mu_1 - \mu_2)F - \mathscr{E}(\mu_1 - \mu_2)\frac{dF}{d\mathscr{E}}\right], \tag{53}$$

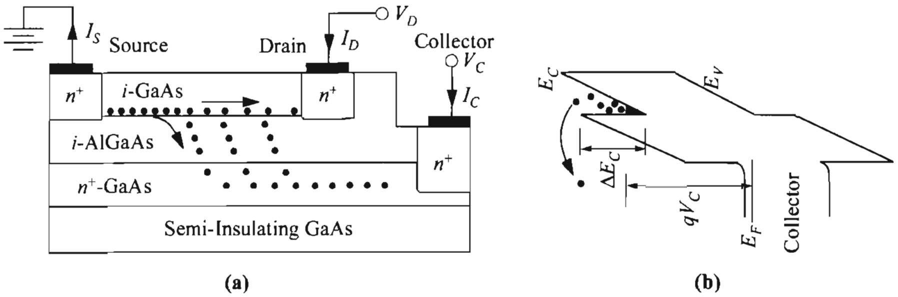
and it can be shown to be negative for a proper choice of μ_1, μ_2, F, and $dF/d\mathscr{E}$. In the GaAs/AlGaAs modulation-doped system, $\mu_1 \approx 8{,}000$ cm^2/V-s and μ_2 is less than 500 cm^2/V-s at room temperature. Experimental data show that the current peak-to-valley ratio is not very high, with a maximum value around 1.5. Computer simulations show that a ratio of more than 2 can be achieved.

One of the advantages of the RST diode is high-speed operation. The response time is limited by the movement of carriers across the heterointerface between the two materials, and is much faster than in a traditional diode where the transit time of carriers between the cathode and anode is the dominating factor.

10.3.2 Real-Space-Transfer Transistor

The *real-space-transfer* (RST) *transistor* is a three-terminal version of a RST diode. In an RST transistor, the third terminal contacts the material of higher conduction band to extract the emitted hot carriers and also to control the transverse field for efficient carrier transfer. The RST transistor was proposed as a *negative-resistance field-effect transistor* (NERFET) by Kastalsky and Luryi in 1983,[52] and was subsequently realized in a GaAs/AlGaAs modulation-doped heterostructure by Kastalsky et al. in 1984.[53]

Typical structure of an RST transistor is shown in Fig. 22, which also indicates the carrier movement and the energy band perpendicular to the heterojunction. As seen, the hot carriers emitted over the barrier are collected by the third terminal and con-

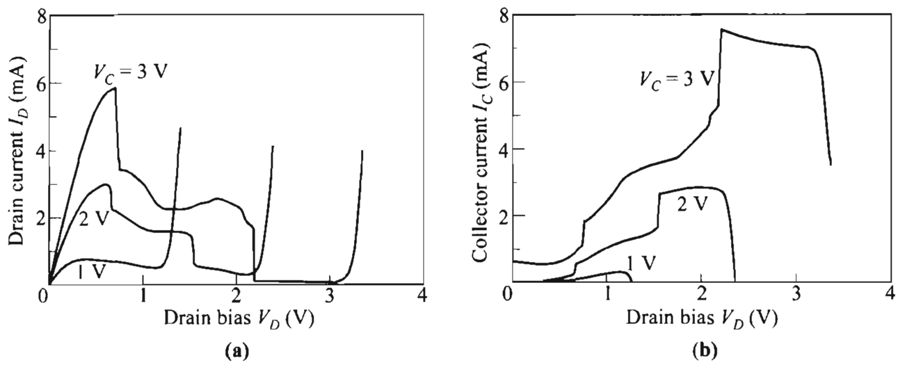


(a) (b)

Fig. 22 (a) Schematic structure of a real-space-transfer transistor and (b) its energy-band diagram perpendicular to the channel.

tribute to another terminal current. For this reason, this third terminal has been called the *collector*. So unlike the RST diode, the mobility in the barrier layer is not relevant. The channel current now decreases because the total carrier density is reduced. This decrease of current is due to *density modulation*, rather than *mobility modulation* as in the RST diode. The current at the source is similar to the channel current of an FET such as a MOSFET or MODFET. The collector is analogous to the gate, which can modulate the channel carrier density and current, with the additional function of depleting the channel hot carriers after they are emitted over the barrier. In this respect, the drop of the channel current results from the increase of collector current. The sum of the drain current and the collector current is thus the same as the total channel current of an insulated-gate FET.

The I-V characteristics for an RST transistor are shown in Fig. 23. At low V_D, the source-drain current is a standard FET current. The collector current is low and its terminal modulates the channel current as an insulated gate. At higher V_D, carriers start to become more and more energetic and begin to spill over the collector barrier. The



(a) (b)

Fig. 23 Terminal currents of a real-space-transfer transistor. (a) Drain current and (b) collector current vs. drain bias. (After Ref. 54.)

collector current is a hot-electron current and it increases with the longitudinal field, or V_D. This current usually occurs in the saturation region in which the field is nonuniform and has a higher local peak near the drain. Kirchhoff's current law requires that

$$I_S = I_D + I_C. \tag{54}$$

The drain current has to drop when I_C rises, giving a negative differential resistance (NDR) dV_D/dI_D. The device is thus a variable NDR device whose characteristics are controlled by the collector terminal. A maximum peak-to-valley ratio of the drain current of more than 340,000 at room temperature has been observed.[55]

We now analyze qualitatively the collector current to gain some understanding of the RST transistor operation. The change of I_D is related to the collector current by

$$\frac{dI_D}{Wdx} = -J_C, \tag{55}$$

where W is the channel width dimension. The collector current is due to thermionic emission of hot electrons. A simple analysis is to assume that the hot carriers have a Maxwellian distribution with a mean electron temperature T_e that is higher than the room or lattice temperature T_o. This thermionic-emission current is given by

$$J_C(x) = qvn(x)\exp\left(-\frac{\Delta E_C}{kT_e}\right) \tag{56}$$

with electron velocity

$$v = \sqrt{\frac{kT_e}{2\pi m^*}}. \tag{57}$$

ΔE_C is the barrier height due to the conduction-band discontinuity shown in Fig. 22b. The electron temperature is related directly to the high local field near the drain, and it has been shown empirically to be proportional to the square of drain bias.[56] Since the drain current is a drift current,

$$I_D(x) = Wdn(x)qv_s \tag{58}$$

where d is the channel thickness. Solving the differential equations of Eqs. 55, 56, and 58 shows that the electron concentration $n(x)$, assuming a uniform field, decays exponentially from source to drain. The total collector current is the integral of Eq. 56 throughout the channel length L. A more rigorous derivation gives the following similar expression from the thermionic-emission theory:[57]

$$I_C = A^{**}T_o^2W\int_0^L\left\{\exp\left[\frac{V_C - V(x)}{kT_e(x)}\right]\exp\left[\frac{-\Delta E_C}{kT_e(x)}\right] - \exp\left[\frac{-\Delta E_C}{kT_o}\right]\right\}dx \tag{59}$$

where A^{**} is the effective Richardson constant (see Chapter 3) and $V(x)$ is the channel potential.

The increase of V_C has the following effects on the device: (1) channel carriers are increased, (2) the field within AlGaAs for efficient collection of hot carriers is increased, and (3) T_e is decreased due to a redistribution of a more uniform field in the channel. These effects have important and sometimes opposite impacts on the hot-electron current. Figure 23 shows three distinct regions according to different values of V_D. At low V_D, carriers do not have enough energy from the longitudinal field to surmount the barrier. The channel is modulated by the collector, and characteristics are similar to the linear region of an FET. The most interesting region is with medium V_D, corresponding to the saturation region of an FET. However, NDR can be observed only for high values of V_C. At low V_C, the transverse field is not high enough for an efficient collection of the carriers within the AlGaAs layer. This is compounded by the space-charge effect, which further reduces the transverse field. The voltage built up by the space-charge effect in the barrier film is given by

$$\Delta V = \frac{J_C l^2}{2\varepsilon_s v_s} \tag{60}$$

where l is the AlGaAs thickness. A quick estimate shows that ΔV can be as high as 2 V.[58] Another interesting feature in this regime is that both positive and negative transconductance (dI_D/dV_C) coexist. Positive transconductance is due to the increase of channel carriers and also the decrease of T_e and I_C with V_C. Negative transconductance is due to the increase of transverse field and I_C with V_C. Finally, the third region is at high V_D, where leakage starts between the drain and the collector.

The intrinsic speed of a RST transistor is limited by two time constants, the energy relaxation time to establish T_e and the time-of-flight within the high-field region near the drain. The latter is transit time over a very short distance, unlike in an FET, where the total distance is from source to drain. Both of these time constants are around 1 ps. Consequently, the RST transistor is proposed to be a very fast device with an ultimate speed, for a well-scaled device, around 100 GHz. A cutoff frequency higher than 70 GHz has been demonstrated.[59]

It has been proposed to use the collector current as the main output current, as *charge-injection transistor* (CHINT), and the drain as the input.[60] There is no difference between a regular RST transistor and the CHINT in structure. Since the I_C is limited by a barrier, its operation is closer to that of a potential-effect transistor (like bipolar transistor) than to an FET. For this reason, the source and drain terminals are analogous to the emitter and base terminals respectively, of a bipolar transistor. The transistor current I_C is determined by the electron temperature T_e created by V_D. Operation of a CHINT can be compared to a vacuum diode whose cathode filament is heated resistively by a current.[60] It can be shown that the transconductance

$$g_m = \frac{dI_C}{dV_D} \tag{61}$$

can be quite high. A maximum g_m of ≈ 1.1 S/mm has been obtained.[59] In the CHINT operation, the NDR is not important.

A three-terminal RST transistor has the following advantages: (1) control of variable NDR, (2) high peak-to-valley current ratio, (3) high-speed operation (since emitted carriers are drained away and will not return to the main channel), (4) high transconductance g_m, and (5) the potential of being a functional device (single device capable of performing functions).

Velocity-Modulation Transistor. The *velocity-modulation transistor* (VMT) was proposed in 1982 as a ultrahigh-speed device.[61] It represents a new class of field-effect transistors in which the source-drain current is not modulated by the amount of carriers or charge induced by the gate (*density modulation*). Rather, the modulation of current is due to a change in carrier velocity (*velocity modulation* or *mobility modulation*), with the unique feature that the total number of carriers remains constant. This can be accomplished by the real-space transfer of channel carriers between two parallel adjacent channels of different mobilities. The main advantage of the velocity-modulation transistor is the intrinsic device speed. For a conventional FET, when a gate bias is applied abruptly to turn the transistor on, the charge induced by the gate comes from the source. By the same token, the off-state requires the charge to dissipate through the drain. The intrinsic speed of a standard FET is thus limited by the transit time between the source and the drain. In the VMT, the change of state is accomplished by transferring charges between two channels that can be much closer than the channel length between the source and drain. Computer simulations show that the response time can be as short as 0.2 ps. So the requirement for a short channel length for speed is removed. Unfortunately, the concept has not yet been demonstrated experimentally. To take full advantage of the VMT, the gate voltage has to be within a range such that the total charge in the channels is conserved. This is not trivial to demonstrate experimentally because similar output characteristics can be achieved by conventional FET action where higher gate voltage induces an extra channel charge. Independent measurements, such as Hall measurement, are required to confirm a VMT action. It should be pointed out that for short-channel devices, the carrier velocity approaches a saturation velocity at high fields. In this operation regime, the difference in mobility is less meaningful than the difference in the saturation velocity of the two channels. This difference in saturation velocity, in practice, is usually much less than that in mobility. This drawback limits the current drive of the device.

REFERENCES

1. J. B. Gunn, "Microwave Oscillation of Current in III-V Semiconductors," *Solid State Commun.*, **1**, 88 (1963).
2. B. Gunn, "Instabilities of Current in III-V Semiconductors," *IBM J. Res. Dev.*, **8**, 141 (1964).
3. H. Kroemer, "Theory of the Gunn Effect," *Proc. IEEE*, **52**, 1736 (1964).
4. B. K. Ridley and T. B. Watkins, "The Possibility of Negative Resistance Effects in Semiconductors," *Proc. Phys. Soc. Lond.*, **78**, 293 (1961).
5. B. K. Ridley, "Anatomy of the Transferred-Electron Effect in III-V Semiconductors," *J. Appl. Phys.*, **48**, 754 (1977).
6. C. Hilsum, "Transferred Electron Amplifiers and Oscillators," *Proc. IRE*, **50**, 185 (1962).

7. C. Hilsum, "Historical Background of Hot Electron Physics," *Solid-State Electron.*, **21**, 5 (1978).
8. A. R. Hutson, A. Jayaraman, A. G. Chynoweth, A. S. Coriell, and W. L. Feldmann, "Mechanism of the Gunn Effect from a Pressure Experiment," *Phys. Rev. Lett.*, **14**, 639 (1965).
9. J. W. Allen, M. Shyam, Y. S. Chen, and G. L. Pearson, "Microwave Oscillations in $GaAs_{1-x}P_x$ Alloys," *Appl. Phys. Lett.*, 7, 78 (1965).
10. J. E. Carroll, *Hot Electron Microwave Generators*, Edward Arnold, London, 1970.
11. P. J. Bulman, G. S. Hobson, and B. S. Taylor, *Transferred Electron Devices*, Academic, New York, 1972.
12. B. G. Bosch and R. W. H. Engelmann, *Gunn-Effect Electronics*, Wiley, New York, 1975.
13. H. W. Thim, "Solid State Microwave Sources," in C. Hilsum, Ed., *Handbook on Semiconductors*, Vol. 4, *Device Physics*, North-Holland, Amsterdam, 1980.
14. M. Shur, *GaAs Devices and Circuits*, Plenum, New York, 1987.
15. D. E. Aspnes, "GaAs Lower Conduction Band Minimum: Ordering and Properties;" *Phys. Rev.*, **14**, 5331 (1976).
16. H. D. Rees and K. W. Gray, "Indium Phosphide: A Semiconductor for Microwave Devices," *Solid State Electron Devices*, **1**, 1 (1976).
17. D. E. McCumber and A. G. Chynoweth, "Theory of Negative Conductance Application and Gunn Instabilities in 'Two-Valley' Semiconductors," *IEEE Trans. Electron Dev.*, **ED-13**, 4 (1966).
18. K. Sakai, T. Ikoma, and Y. Adachi, "Velocity-Field Characteristics of $Ga_xIn_{1-x}Sb$ Calculated by the Monte Carlo Method," *Electron. Lett.*, **10**, 402 (1974).
19. R. E. Hayes and R. M. Raymond, "Observation of the Transferred-Electron Effect in GaInAsP," *Appl. Phys. Lett.*, **31**, 300 (1977).
20. J. R. Hauser, T. H. Glisson, and M. A. Littlejohn, "Negative Resistance and Peak Velocity in the Central (000) Valley of III-V Semiconductors," *Solid-State Electron.*, **22**,'487 (1979).
21. J. G. Ruch and G. S. Kino, "Measurement of the Velocity-Field Characteristics of Gallium Arsenide," *Appl. Phys. Lett.*, **10**, 40 (1967).
22. P. N. Butcher and W. Fawcett, "Calculation of the Velocity-Field Characteristics for Gallium Arsenide," *Phys. Lett.*, **21**, 489 (1966).
23. M. A. Littlejohn, J. R. Hauser, and T. H. Glisson, "Velocity-Field Characteristics of GaAs with Γ-L-X Conduction-Band Ordering," *J. Appl. Phys.*, **48**, 4587 (1977).
24. I. Mojzes, B. Podor, and I. Balogh, "On the Temperature Dependence of Peak Electron Velocity and Threshold Field Measured on GaAs Gunn Diodes," *Phys. Status Solidi*, **39**, K123 (1977).
25. P. N. Butcher, "Theory of Stable Domain Propagation in the Gunn Effect," *Phys. Lett.*, **19**, 546 (1965).
26. P. N. Butcher, W. Fawcett, and C. Hilsum, "A Simple Analysis of Stable Domain Propagation in the Gunn Effect," *Br. J. Appl. Phys.*, **17**, 841 (1966).
27. J. A. Copeland, "Electrostatic Domains in Two-Valley Semiconductors," *IEEE Trans. Electron Dev.*, **ED-13**, 187 (1966).
28. M. Shaw, H. L. Grubin, and P. R. Solomon, *The Gunn-Hilsum Effect*, Academic, New York, 1979.

29. G. S. Kino and I. Kuru, "High-Efficiency Operation of a Gunn Oscillator in the Domain Mode," *IEEE Trans. Electron Dev.*, **ED-16**, 735 (1969).
30. H. W. Thim, "Computer Study of Bulk GaAs Devices with Random One-Dimensional Doping Fluctuations," *J. Appl. Phys.*, **39**, 3897 (1968).
31. J. A. Copeland, "A New Mode of Operation for Bulk Negative Resistance Oscillators," *Proc. IEEE*, **54**, 1479 (1966).
32. J. A. Copeland, "LSA Oscillator Diode Theory," *J. Appl. Phys.*, **38**, 3096 (1967).
33. M. R. Barber, "High Power Quenched Gunn Oscillators," *Proc. IEEE*, **56**, 752 (1968).
34. H. W. Thim and M. R. Barber, "Observation of Multiple High-Field Domains in *n*-GaAs," *Proc. IEEE*, **56**, 110 (1968).
35. G. S. Hobson, *The Gunn Effect*, Clarendon, Oxford, 1974.
36. H. W. Thim and W. Haydl, "Microwave Amplifier Circuit Consideration," in M. J. Howes and D. V. Morgan, Eds., *Microwave Devices*, Wiley, New York, 1976, Chap. 6.
37. D. Jones and H. D. Rees, "Electron-Relaxation Effects in Transferred-Electron Devices Revealed by New Simulation Method," *Electron. Lett.*, **8**, 363 (1972).
38. H. Kroemer, "Hot Electron Relaxation Effects in Devices," *Solid-State Electron.*, **21**, 61 (1978).
39. H. Kroemer, "The Gunn Effect under Imperfect Cathode Boundary Condition," *IEEE Trans. Electron Dev.*, **ED-15**, 819 (1968).
40. M. M. Atalla and J. L. Moll, "Emitter Controlled Negative Resistance in GaAs," *Solid-State Electron.*, **12**, 619 (1969).
41. S. P. Yu, W. Tantraporn, and J. D. Young, "Transit-Time Negative Conductance in GaAs Bulk-Effect Diodes," *IEEE Trans. Electron Dev.*, **ED-18**, 88 (1971).
42. D. J. Colliver, L. D. Irving, J. E. Pattison, and H. D. Rees, "High-Efficiency InP Transferred-Electron Oscillators," *Electron. Lett.*, **10**, 221 (1974).
43. K. W. Gray, J. E. Pattison, J. E. Rees, B. A. Prew, R. C. Clarke, and L. D. Irving, "InP Microwave Oscillator with 2-Zone Cathodes," *Electron. Lett.*, **11**, 402 (1975).
44. H. D. Rees, "Time Response of the High-Field Electron Distribution Function in GaAs," *IBM J. Res. Dev.*, **13**, 537 (1969).
45. H. Eisele and R. Kamoua, "Submillimeter-Wave InP Gunn Devices," *IEEE Trans. Microwave Theory Tech.*, **52**, 2371 (2004).
46. A. Ataman and W. Harth, "Intrinsic FM Noise of Gunn Oscillators," *IEEE Trans. Electron Dev.*, **ED-20**, 12 (1973).
47. M. Shoji, "Functional Bulk Semiconductor Oscillators," *IEEE Trans. Electron Dev.*, **ED-14**, 535 (1967).
48. Z. S. Gribnikov, "Negative Differential Conductivity in a Multilayer Heterostructure," *Soviet Phys.–Semiconductors*, **6**, 1204 (1973). Translated from *Fizika i Teknika Poluprovodnikov*, **6**, 1380 (1972).
49. K. Hess, H. Morkoc, H. Shichijo, and B. G. Streetman, "Negative Differential Resistance Through Real-Space Electron Transfer," *Appl. Phys. Lett.*, **35**, 469 (1979).
50. M. Keever, H. Shichijo, K. Hess, S. Banerjee, L. Witkowski, H. Morkoc, and B. G. Streetman, "Measurements of Hot-Electron Conduction and Real-Space Transfer in GaAs-$Al_xGa_{1-x}As$ Heterojunction Layers," *Appl. Phys. Lett.*, **38**, 36 (1981).

51. Z. S. Gribnikov, K. Hess, and G. A. Kosinovsky, "Nonlocal and Nonlinear Transport in Semiconductors: Real-Space Transfer Effects," *J. Appl. Phys.*, **77**, 1337 (1995).
52. A. Kastalsky and S. Luryi, "Novel Real-Space Hot-Electron Transfer Devices," *IEEE Electron Dev. Lett.*, **EDL-4**, 334 (1983).
53. A. Kastalsky, S. Luryi, A. C. Gossard, and R. Hendel, "A Field-Effect Transistor with a Negative Differential Resistance," *IEEE Electron Dev. Lett.*, **EDL-5**, 57 (1984).
54. P. M. Mensz, S. Luryi, A. Y. Cho, D. L. Sivco, and F. Ren, "Real-Space Transfer in Three-Terminal InGaAs/InAlAs/InGaAs Heterostructure Devices," *Appl. Phys. Lett.*, **56**, 2563 (1990).
55. C. L. Wu, W. C. Hsu, H. M. Shieh, and M. S. Tsai, "A Novel δ-Doped GaAs/InGaAs Real-Space Transfer Transistor with High Peak-to-Valley Ratio and High Current Driving Capability," *IEEE Electron Dev. Lett.*, **EDL-16**, 112 (1995).
56. S. Luryi, "Hot-Electron transistors," in S. M. Sze, Ed., *High-Speed Semiconductor Devices,* Wiley, New York, 1990.
57. E. J. Martinez, M. S. Shur, and F. L. Schuermeyer, "Gate Current Model for the Hot-Electron Regime of Operation in Heterostructure Field Effect Transistors," *IEEE Trans. Electron Dev.*, **ED-45**, 2108 (1998).
58. S. Luryi and A. Kastalsky, "Hot Electron Injection Devices," *Superlattices and Microstructures*, **1**, 389 (1985).
59. G. L. Belenky, P. A. Garbinski, P. R. Smith, S. Luryi, A. Y. Cho, R. A. Hamm, and D. L. Sivco, "Microwave Performance of Top-Collector Charge Injection Transistors on InP Substrates," *Semicond. Sci. Technol.*, **9**, 1215 (1994).
60. S. Luryi, A. Kastalsky, A. C. Gossard, and R. H. Hendel, "Charge Injection Transistor Based on Real-Space Hot-Electron Transfer," *IEEE Trans. Electron Dev.*, **ED-31**, 832 (1984).
61. H. Sakaki, "Velocity-Modulation Transistor (VMT)–A New Field-Effect Transistor Concept," *Jpn. J. Appl. Phys.*, **21**, L381 (1982).

PROBLEMS

1. An InP TED is 0.5 μm long with a cross sectional area of 10^{-4} cm^2 and is operated in the transit-time mode.
   (a) Find the minimum electron density n_0 required for transit-time mode.
   (b) Find the time between current pulses.
   (c) Calculate the power dissipated in the device, if it is biased at one half the threshold.
2. For an InP TED operated in the transit-time dipole-layer mode, the device length is 20 μm, and the doping is $n_0 = 10^{15}$ cm^{-3}. If the current density is 3.2 kA/cm^2 when the dipole is not in contact with the electrodes, find the domain excess voltage, assuming triangular domain.
3. (a) Find the effective density of states in the upper valley N_{CU} of the GaAs conduction band. The upper-valley effective mass is $1.2m_0$.
   (b) The ratio of electron concentrations between the upper and lower valleys is given by $(N_{CU}/N_{CL})\exp(-\Delta E/kT_e)$, where N_{CL} is the effective density of states in the lower valley, $\Delta E = 0.31$ eV is the energy difference, and T_e is the effective electron temperature. Find the ratio at $T_e = 100$ K.

(c) When electrons gain kinetic energies from the electric field, T_e increases. Find the concentration ratio for $T_e = 1500$ K.

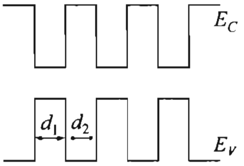
4. In a transferred-electron device, if a domain is suddenly quenched during transit so that the excess domain voltage is changed from V_{ex} to zero in a time that is short compared to the transit time, the change in total current through the device during this time, integrated with respect to time, should give a measure of the charge stored in the domain, Q_0. Relate this charge Q_0 to the domain excess voltage V_{ex} for a triangle electric field distribution, i.e., the field increases linearly from $\mathscr{E}_r$ to $\mathscr{E}_{dom}$ over a distance x_A for the accumulation layer and decreases linearly from $\mathscr{E}_{dom}$ to $\mathscr{E}_r$ over a distance x_D for the depletion layer (assuming that the charge in each layer is uniform).

5. Consider a simple model of RST that neglects the electric field associated with transferred electrons. Assume a periodic multilayer structure as shown with narrow-gap layers of thickness d_1 and wide-gap layers of thickness d_2. The effective masses and mobilities in the layers are m_1 and m_2, and μ_1 and μ_2 respectively, with $\mu_1 > \mu_2$. Take the rate of energy loss to the lattice proportional to $(T_e - T)/\tau$ per electron, where τ is the same for both layers. Further, assume that the effective electron temperature T_e is the same for both layers, so that no energy is transferred on average as electrons jump between layers. The total density of electrons is fixed, $n = n_1 + n_2$ is constant.

(a) Derive the energy balance equation.

(b) Express the ratio n_1/n_2 in terms of T_e and the barrier height ϕ.

(c) Derive the current-field characteristics in a parametric form $\mathscr{E} = \mathscr{E}(T_e)$, $J = J(T_e)$ and plot $J(\mathscr{E})$ for a few values of the parameters. What is necessary to achieve high NDR in the source-drain current-field characteristic?



6. In a RST device, the collector current can be expressed as $I_c = A\exp(-\phi/kT_e)$, where A is a constant, ϕ the barrier height, and T_e the electron temperature. Assume $(T_e - T)/T = BV_{SD}^m$, where T is the lattice temperate, B and m are constants, and V_{SD} is the applied voltage between source and drain.

(a) Show that

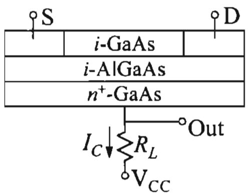
$$f \equiv \left(V_{SD}\frac{d\ \ln J_c}{d\ \ln V_{SD}}\right)^{-1} = \frac{kT_e}{m\phi}\left(\frac{T_e}{T_e - T}\right).$$

(b) Show that when $T_e \gg T$, a plat of f versus V_{SD}^m is a straight line.

7. A real-space-transfer device has hot-electron injection curves shown in Figs. 21 and 22 (pp. 439–440, *High Speed Semiconductor Devices*. Sze, Ed., Wiley, 1990). Estimate the barrier height ϕ in eV for $V_{sub} = 0.25$ V and $V_{SD} = -0.448$ V.

8. A real-space-transfer transistor (Fig. 22) has an intrinsic layer (*i*-AlGaAs) of 0.1 μm in thickness. If the energy relaxation time is 1 ps, and the transit velocity of carriers across the *i*-region is 10^7 cm/s, estimate the cutoff frequency of the device.

9. For a CHINT logic shown, prepare a truth table and show that this a an exclusive NOR logic, i.e., the output voltage will be high only when both source and drain are at the same voltage level.



10. A conventional CHINT has an *n*-type emitter channel and an *n*-type collector (2nd conducting layer). By replacing the *n*-type collector with a *p*-type collector, we can realize a novel light emitting device. We can construct such a device with the following lattice-matched layers: 1 μm *p*-type AlInAs (3×10^{18} cm^{-3} with $E_g = 1.5$ eV), 50 nm *p*-type GaInAs (10^{17}), 200 nm undoped AlInAs barrier, 50 nm *n*-type GaInAs (10^{17}), and 2.5 nm AlInAs *n*-type (10^{19}). Draw the band diagram of the structure from the top 2.5 nm AlInAs layer to the bottom 1 μm AlInAs layer at thermal equilibrium ($\Delta E_C = 0.53$ eV, $\Delta E_V = 0.22$ eV). Mark all bandgap energies and layer thicknesses on the diagram. Find the wavelength of the emitted light.

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

11

Thyristors and Power Devices

11.1 INTRODUCTION
11.2 THYRISTOR CHARACTERISTICS
11.3 THYRISTOR VARIATIONS
11.4 OTHER POWER DEVICES

11.1 INTRODUCTION

Power semiconductor devices can be generally grouped into two functions. The first as a *switch* is to control the power delivered to the load. In this case, only the two extreme states are critical. The *on*-state ideally should be a short whereas the *off*-state an open. The second function as a power *amplifier* is to amplify an ac signal. For this function, the current gain (in a potential-effect transistor) or the transconductance (in a field-effect transistor) is important. For power applications, both types of devices are required to sustain high voltages and high currents. A good example for a switch is a thyristor which possesses S-type negative differential resistance. Because of the snap-back action and its associated nonlinear effects, the thyristor cannot be used as a power amplifier. On the other hand, a power transistor with smooth I-V characteristics can also be used as a switch. Most devices discussed in this chapter are related to thyristors, and they function only as efficient switching devices. The only devices that can be used for both functions are insulated-gate bipolar transistor (IGBT) and static-induction devices discussed near the end of the chapter.

This chapter only covers power devices whose working principles are different from previous chapters. Power devices based on transistors such as MOSFET, JFET, MESFET, MODFET, bipolar transistor, etc., are commonly used in power applications. Their device principles, however, do not require additional treatment. The main differences for those devices in power applications mainly lie in their structures, which usually have larger dimensions, better heat sinks, and sometimes made on different semiconductor materials.

More-suitable semiconductor materials for power devices are compared in Table 1, which lists their most-important parameters. A high-bandgap material usually has low ionization coefficients, giving rise to low impact ionization and high breakdown voltage. Mobility and saturation velocity are for speed consideration. High thermal conductivity enhances heat conduction and increases the power level. Of the materials shown, SiC and GaN have the best combination. The only drawback is their immature technologies in that the materials are not very reliable or reproducible, and the cost is very high.

11.2 THYRISTOR CHARACTERISTICS

The name *thyristor* applies to a general family of semiconductor devices that exhibit bistable characteristics and can be switched between a high-impedance, low-current *off*-state and a low-impedance, high-current *on*-state. Also, the operations of thyristors are intimately related to the bipolar-transistor action in which both electrons and holes interact with each other in the transport processes. The name *thyristor* is derived from vacuum tube *gas thyratron*, since the electrical characteristics are similar in many respects.

Following Shockley's concept of the *hook collector* in 1950,[1] Ebers developed a two-transistor analogue to explain the characteristics of a basic thyristor, a multilayered *p-n-p-n* device.[2] The detailed device principles and the first working two-terminal *p-n-p-n* devices were reported by Moll et al. in 1956.[3] This work has since served as the basis in the study of thyristors. Subsequently, the control of switching using a third terminal was examined by Mackintosh,[4] and by Aldrich and Holonyak[5] in 1958. Because of their two stable states (*on* and *off*) and low power dissipations in these states, thyristors have found unique usefulness in applications ranging from speed control in home appliances to switching and power conversion in high-voltage transmission lines. Thyristors are now available with current ratings from a few mA to over 5 kA and voltage ratings extending above 10 kV. Comprehensive treatments on the operation and fabrication technology of thyristor can be found in Refs. 6–10.

Table 1 Comparison of Semiconductor Materials for Power Applications

Property	Si	GaAs	SiC(4H)	GaN
Bandgap (eV)	1.12	1.42	3.0	3.4
Dielectric constant	11.9	12.9	10	10.4
Breakdown field (V/cm)	$\approx 3\times10^5$	$\approx 4\times10^5$	$\approx 4\times10^6$	$\approx 4\times10^6$
Saturation velocity (cm/s)	1×10^7	0.7×10^7	2×10^7	1.5×10^7
Peak velocity (cm/s)	1×10^7	2×10^7	2×10^7	$>2\times10^7$
Electron mobility (cm^2/V-s)	1350	6000	800	1000*
Thermal conductivity (W/cm-°C)	1.5	0.46	4.9	1.7

* Modulation-doped channel.

The basic thyristor structure is shown schematically in Fig. 1a. It is a four-layer *p-n-p-n* device having three *p-n* junctions, J1, J2, and J3, in series. Its typical doping profile is shown in Fig. 1b. Note that the *n*1-layer (*n*-base) is much wider than other regions, and it has the lowest doping level for high breakdown voltage. The contact electrode to the outer *p*-layer is called the anode and that to the outer *n*-layer is called the cathode. The gate electrode, also referred to as the base, is connected to the inner *p*-layer (*p*-base). The three-terminal device is commonly called the semiconductor-controlled rectifier (SCR) or simply the thyristor. Without the gate electrode, the device is operated as a two-terminal *p-n-p-n* switch, or Shockley diode.

The basic current-voltage characteristics of a thyristor that have a number of complex regions are shown in Fig. 2. In region 0–1 the device is in the forward blocking or *off*-state with very high impedance. Forward breakover (or switching) occurs at $dV_{AK}/dI_A = 0$, where we define a forward breakover voltage V_{BF} and a switching current I_s (also called the turn-on current). Region 1–2 is the negative-resistance region, and region 2–3 is the forward-conduction mode or *on*-state. At point-2, where again $dV_{AK}/dI_A = 0$, we define the holding current I_h and holding voltage V_h. Region 0–4 is in the reverse-blocking state, and region 4–5 is the reverse-breakdown region. Note that the switching voltage V_{BF} can be varied by the gate current I_g. For a two-terminal Shockley diode, the characteristics are simply equivalent to that of an open-gate, or $I_g = 0$.

A thyristor operated in the forward region is, thus, a bistable device that can switch from a high-impedance, low-current *off*-state to a low-impedance, high-current *on*-state, or vice versa. In this section, we consider the basic thyristor characteristics shown in Fig. 2, that is, the reverse-blocking, forward-blocking, and forward-conduction modes.

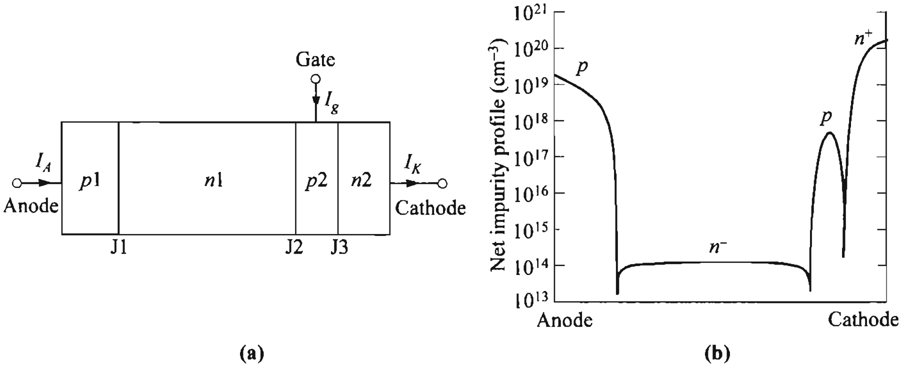


Fig. 1 (a) Schematic structure of thyristor. There exist three *p-n* junctions in series J1, J2, and J3. (b) Typical doping profile.

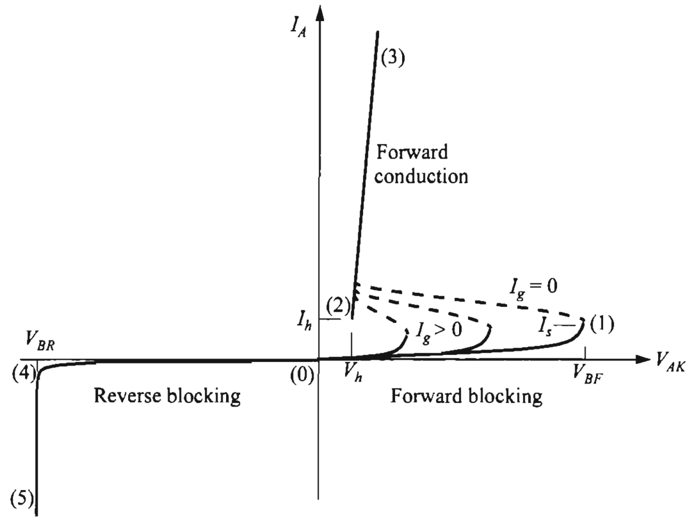


Fig. 2 Current-voltage characteristics of a thyristor. The switching voltage V_{BF} can be lowered by I_g.

11.2.1 Reverse Blocking

Two basic factors limiting the reverse breakdown voltage and the forward breakover voltage are avalanche breakdown and depletion-layer punch-through. In the reverse-blocking mode, the applied anode voltage is negative with respect to the cathode; junctions J1 and J3 are reverse biased and J2 is forward biased. From the doping profile shown (Fig. 1b), most of the applied reverse voltage will drop across J1 and the $n1$-region (Fig. 3a). Depending on the thickness of the $n1$-layer W_{n1}, the breakdown will be caused by avalanche multiplication if the depletion-layer width at breakdown is less than W_{n1}, or caused by punch-through if the whole width W_{n1} is consumed first by the depletion layer, at which point the junction J1 is effectively shorted to J2.

For a one-sided abrupt silicon p^+-n junction with a heavily doped $p1$-region, the avalanche breakdown voltage at room temperature is given by[8,11] (Eq. 104 on p. 110)

$$V_B \approx 6.0 \times 10^{13} (N_{n1})^{-0.75} \tag{1}$$

where N_{n1} is the doping concentration of the $n1$-region in cm^{-3}. The punch-through voltage for the one-sided abrupt junction is given by

$$V_{PT} = \frac{qN_{n1}W_{n1}^2}{2\varepsilon_s}. \tag{2}$$

Figure 3b shows the fundamental limit of the reverse-blocking capability of silicon thyristors.[12] For example, for $W_{n1} = 160$ μm the maximum breakdown voltage is limited to ≈ 1 kV, which occurs at $N_{n1} \approx 8 \times 10^{13}$ cm^{-3}. For lower dopings the break-

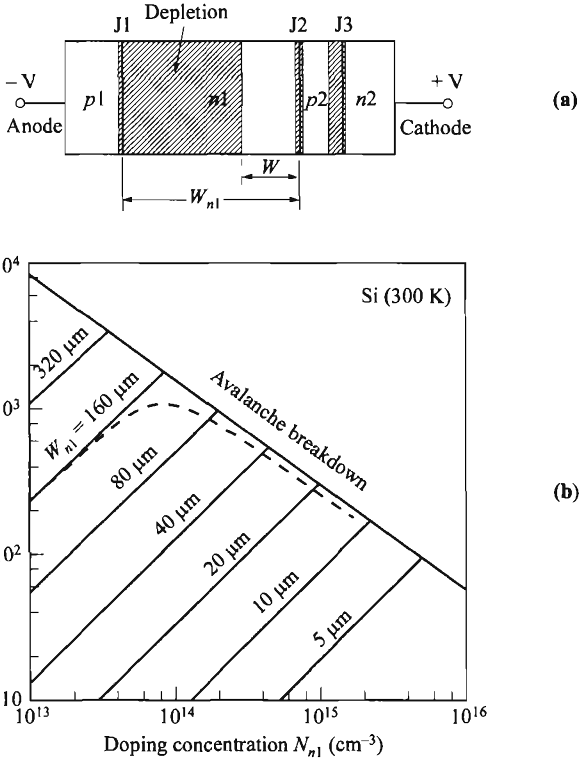


Fig. 3 Reverse-blocking capability of thyristor. (a) Dimensions including depletion width under reverse bias. (b) Reverse-blocking voltage bounded by avalanche (top) and punch-through (parallel lines). W_{n1} and N_{n1} are width and doping of the $n1$-layer. Dashed line is an example for W_{n1} = 160 µm.

down voltage is limited by punch-through and for higher dopings by avalanche multiplication.

The actual reverse-blocking voltage bounded by avalanche should lie below that given by a simple *p-n* junction because of the *p-n-p* bipolar transistor current gain. This is analogous to the analysis of breakdown in a bipolar transistor. The reverse-breakdown condition corresponds to that for the common-emitter configuration, which is $M = 1/\alpha_1$ (Eq. 45 on p. 257) where M is the avalanche multiplication factor and α_1 the *p-n-p* bipolar common-base current gain. The breakdown voltage is given by

$$V_{BR} = V_B(1-\alpha_1)^{1/n} \tag{3}$$

where V_B is the avalanche breakdown voltage of the J1 junction, and n a constant (≈ 6 for Si p^+-n diodes). Since $(1-\alpha_1)^{1/n}$ is less than unity, the reverse breakdown voltage of a thyristor will be less than V_B. We can further estimate the influence of α_1 on V_{BR}.

The injection efficiency γ is close to unity for most practical situations, since the $p2$-region (emitter) is heavily doped. The current gain is, therefore, reduced to the transport factor α_T:

$$\alpha_1 = \gamma\alpha_T \approx \alpha_T \approx \operatorname{sech}\left(\frac{W}{L_{n1}}\right) \tag{4}$$

where L_{n1} is the hole diffusion length in the $n1$-region and W is the neutral $n1$-region

$$W \approx W_{n1}\left[1 - \sqrt{\frac{V_{AK}}{V_{PT}}}\right]. \tag{5}$$

For a given W_{n1} and L_{n1}, the ratio W/W_{n1} will decrease as the reverse voltage increases. Therefore, the base transport factor α_T becomes more important as the reverse voltage approaches the punch-through limit. Figure 3 shows an example for the reverse-blocking voltage with W_{n1} = 160 μm and L_{n1} = 150 μm (dashed curve). Note that V_{BR} approaches the V_{PT} for low dopings in the $n1$-region. As the doping increases, V_{BR} always lies slightly below V_B, because of the finite value of W/L_{n1}.

Neutron-Transmutation Doping. For high-power, high-voltage thyristors, large areas are used, frequently an entire wafer (up to 100 mm or larger in diameter) for a single device. This size imposes stringent requirements on the uniformity of the starting material. To obtain tight tolerance for the resistivity and homogeneous distribution of impurity dopant, the technique of neutron-transmutation doping is employed.[13] Usually the float-zone silicon wafer having an average resistivity well in excess of that required is used to start. The wafer is then irradiated with thermal neutrons and the process first produces a heavier silicon isotope which is unstable. This silicon isotope changes to a new element one atomic number higher—in this case phosphorus, an n-type dopant for silicon. The neutron-transmutation process is represented by:

$$\mathrm{Si}_{14}^{30} + \text{neutron} \rightarrow \mathrm{Si}_{14}^{31} \underset{2.62\ \mathrm{h}}{\rightarrow} \mathrm{P}_{15}^{31} + \beta \text{ ray}. \tag{6}$$

The second reaction emits beta particles and has a decay half-life of 2.62 h. Since the penetration range of neutron irradiation in silicon is very large, about 100 cm, doping is very uniform throughout the wafer thickness. Figure 4 compares the lateral macroscopic resistivity distributions in conventionally doped silicon and in silicon doped by neutron transmutation, obtained by spreading resistance measurements. The resistivity variations are about ±15% for the conventionally doped silicon and about ±1% for neutron-transmutation doped.

Beveled Structures. To maximize the breakdown voltage in a thyristor, usually planar junctions formed by diffusion or implantation are used, since cylindrical or spherical junctions have lower breakdown voltages (refer to Section 2.4.3). Even for planar junctions, premature breakdown can still occur at the edge where the junction terminates. By using properly beveled structures, the surface field can be lowered significantly compared to the bulk value, ensuring that the breakdown will occur uniformly in the bulk.

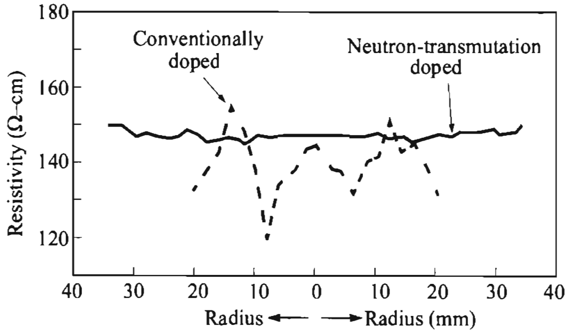


Fig. 4 Comparison of doping uniform of high-resistivity Si by conventional doping and neutron-transmutation doping. (After Ref. 14.)

Figure 5 shows the beveled structures. A positive bevel angle is defined as in a junction of decreasing cross-sectional area when going from the heavily doped side to the lightly doped side (Fig. 5a). The negative bevel angle, on the other hand, has an increasing area going in the same direction (Fig. 5b). Two beveled thyristor structures are shown in Figs. 5c and d. Figure 5c has a negative bevel angle for junctions J2 and J3, and a positive bevel angle for J1. Figure 5d has positive bevel angles for all three junctions.[15]

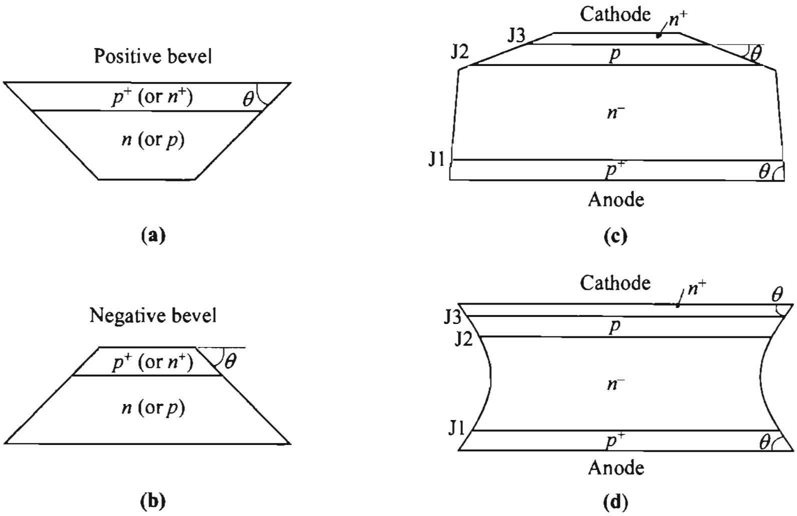

Positive bevel
p^+ (or n^+)
θ
n (or p)
(a)
Negative bevel
p^+ (or n^+)
θ
n (or p)
(b)
Cathode
n^+
J3
J2
p
θ
n^-
J1
p^+
θ
Anode
(c)
Cathode
n^+
θ
J3
J2
p
n^-
J1
p^+
θ
Anode
(d)

Fig. 5 p-n junctions with (a) positive bevel angle and (b) negative bevel angle. (c) Thyristor with two negative bevel angles (J2, J3) and one positive bevel angle (J1). (d) Thyristor with three positive bevel angles.

For the positive-beveled junction, the surface field, as a first-order approximation, is reduced by the factor $\sin\theta$. Figure 6 shows the calculated values of the electric field from a two-dimensional Poisson equation for a p^+-n junction under a reverse bias of 600 V. The internal electric field in the bulk is also shown. Note that the peak field on the bevel surface is always less than that in the bulk, and as the bevel angle is reduced, so is the peak electric field. The position of the peak field also shifts into the lightly doped region as the bevel angle is reduced. The breakdown voltage for the positive-beveled junction is dominated by the internal junction since the edge effect does not cause premature breakdown.

For the negative-beveled junction, the trend is more complicated and is not monotonic. Calculated results for the peak surface field as a function of the negative bevel angle are shown in Figure 7. It is seen here that for most of the angles, the peak field at the bevel edge is higher compared to the internal junction. However, if the negative bevel angle is small enough, the peak field starts to drop again. In order to have the surface peak field smaller than the internal field, the negative bevel angle has to be smaller than ≈ 20°.

In summary, to avoid breakdown initiated by the edge effect, the bevel angle should be either positive, or negative but less than ≈ 20°. Returning to the thyristor structures, Fig. 5c is a common design where J2 and J3 have small negative bevel angles. Structure in Fig. 5d is more ideal since all three junctions have positive bevel angles. However, it is more difficult to fabricate and is, thus, much less common.

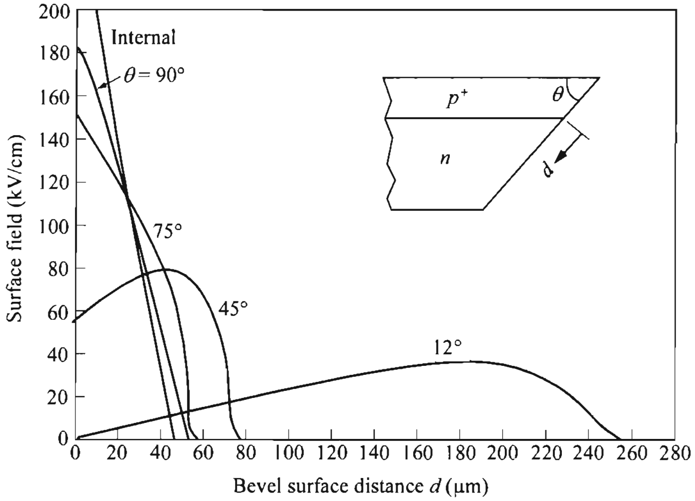


Fig. 6 Surface fields along bevel with positive bevel angles. (After Ref. 16.)

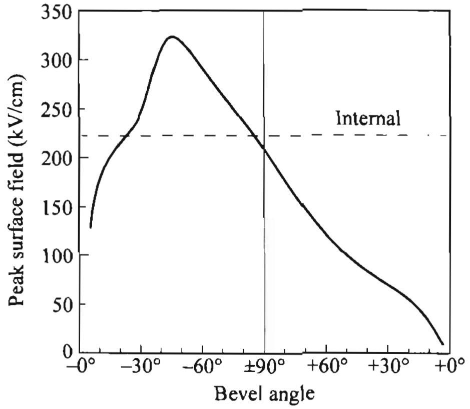


Fig. 7 Peak electric field along bevel surface for both positive and negative bevel angles. (After Ref. 16.)

11.2.2 Forward Blocking

For forward blocking, the anode voltage is positive with respect to the cathode, and only the center junction J2 is reverse biased. Junctions J1 and J3 are forward biased. Most of the applied voltage will drop across J2 ($V_{AK} \approx V_2$). To understand the forward-blocking characteristics, we shall use the method of the two-transistor analog.[2] The thyristor can be considered as a *p-n-p* transistor and an *n-p-n* transistor connected with the collector of one transistor attached to the base of the other, and vice versa, as shown in Fig. 8. The center junction J2 acts as the collector of holes from J1 and of electrons from J3.

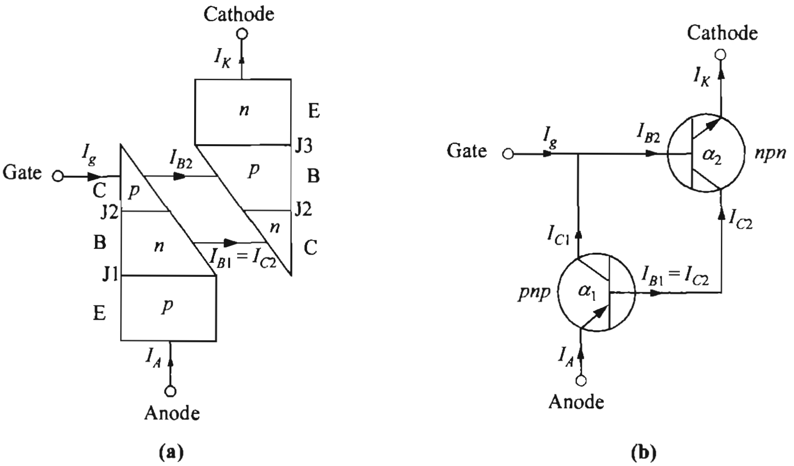

Cathode
I_K
n
E
J3
B
J2
C
p
n
I_{B2}
$I_{B1} = I_{C2}$
Gate
I_g
C
p
J2
B
n
J1
E
p
I_A
Anode
(a)
Cathode
I_K
Gate
I_g
I_{B2}
α_2
npn
I_{C2}
I_{C1}
pnp
α_1
$I_{B1} = I_{C2}$
I_A
Anode
(b)

Fig. 8 (a) Two-transistor approximation of a three-terminal thyristor, bisected into (but connected) *p-n-p* and *n-p-n* bipolar transistors. (b) Its circuit representation using transistor notations.

The relationship between emitter, collector, and base currents (I_E, I_C, and I_B, respectively) and the dc common-base current gain α for a *p-n-p* transistor is given by

$$I_C = \alpha I_E + I_{CO}, \tag{7}$$

$$I_E = I_C + I_B, \tag{8}$$

where I_{CO} is the collector-base reverse saturation current. Similar relationships can be obtained for an *n-p-n* transistor, except that the currents are reversed. From Fig. 8b it is evident that the collector current of the *n-p-n* transistor provides the base drive for the *p-n-p* transistor. Also, the collector current of the *p-n-p* transistor along with gate current I_g supplies the base drive for the *n-p-n* transistor. Thus a regeneration situation results when the total loop gain exceeds unity.

The base current of the *p-n-p* transistor is

$$I_{B1} = (1 - \alpha_1)I_A - I_{CO1} \tag{9}$$

which is supplied by the collector of the *n-p-n* transistor. The collector current of the *n-p-n* transistor with a dc common-base current gain α_2 is given by

$$I_{C2} = \alpha_2 I_K + I_{CO2}. \tag{10}$$

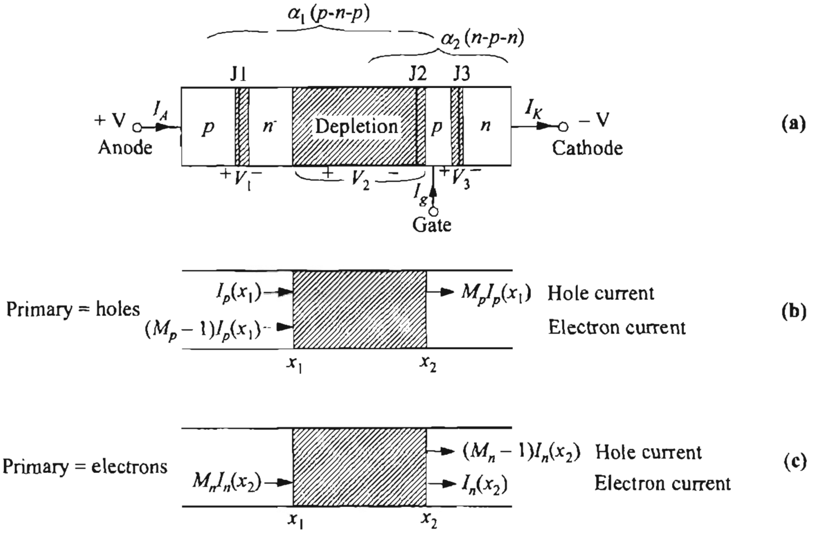
By equating I_{B1} and I_{C2}, and since $I_K = I_A + I_g$, we obtain

$$I_A = \frac{\alpha_2 I_g + I_{CO1} + I_{CO2}}{1 - (\alpha_1 + \alpha_2)} \qquad (\alpha_1 + \alpha_2) < 1\,. \tag{11}$$

It will be shown later that both α_1 and α_2 are function of the current I_A and generally increase with increasing current. Equation 11 gives the static characteristic of the device up to the breakover voltage. Beyond this point the device acts as a *p-i-n* diode. Note that all the current components in the numerator of Eq. 11 are small, hence I_A is small unless $(\alpha_1 + \alpha_2)$ approaches unity. At that point the denominator of the equation approaches zero and forward breakover or switching will occur.

Forward Breakover Voltage. Equation 11, in the first order, gives a constant current independent of V_{AK}. If V_{AK} continues to increase, not only will α_1 and α_2 increase towards the condition of $(\alpha_1 + \alpha_2) = 1$, the high field also initiates carrier multiplication. The interaction of gain and multiplication will decide the switching condition and the breakover voltage V_{BF}. To obtain V_{BF}, we shall consider a general thyristor as shown in Fig. 9. Reference directions for voltages and currents are shown in the figure. We assume that the center junction J2 of the device remains reverse biased. We also assume that the voltage drop V_2 across this junction is sufficient to produce avalanche multiplication of carriers as they travel across the depletion region. We denote the multiplication factor for electrons by M_n and that for holes by M_p; both are functions of V_2. Because of the multiplication, a steady hole current $I_p(x_1)$ entering the depletion region at x_1 becomes $M_p I_p(x_1)$ at $x = x_2$. A similar result will be obtained for an electron current $I_n(x_2)$ entering the depletion layer at x_2. The total current I crossing J2 is given by

$$I = M_p I_p(x_1) + M_n I_n(x_2). \tag{12}$$



α₁ (p-n-p)
α₂ (n-p-n)
J1
J2
J3
+V
Anode
I_A
p
n⁻
Depletion
p
n
I_K
−V
Cathode
+V₁−
+ V₂ −
+V₃−
I_g
Gate
(a)
Primary = holes
I_p(x₁)
(M_p − 1)I_p(x₁)
M_pI_p(x₁)
Hole current
Electron current
x₁
x₂
(b)
Primary = electrons
M_nI_n(x₂)
(M_n − 1)I_n(x₂)
I_n(x₂)
Hole current
Electron current
x₁
x₂
(c)

Fig. 9 Thyristor under high forward bias. (a) Avalanche multiplication occurs in the depletion layer of junction J2 which is under reverse bias. In (b) and (c), currents of the opposite type of carriers from the primary are also generated.

Since $I_p(x_1)$ is actually the collector current of the *p-n-p* transistor, we can express $I_p(x_1)$ as in Eq. 7,

$$I_p(x_1) = \alpha_1(I_A)I_A + I_{CO1}. \tag{13a}$$

Similarly, we can express the primary electron current $I_n(x_2)$ as

$$I_n(x_2) = \alpha_2(I_K)I_K + I_{CO2}. \tag{13b}$$

Substituting Eqs. 13a and 13b into Eq. 12 yields

$$I = M_p[\alpha_1(I_A)I_A + I_{CO1}] + M_n[\alpha_2(I_K)I_K + I_{CO2}]. \tag{14}$$

If we assume that $M_p = M_n = M$ which is a function of V_2, Eq. 14 reduces to

$$I = M(V_2)[\alpha_1(I_A)I_A + \alpha_2(I_K)I_K + I_0] \tag{15}$$

where $I_0 = I_{CO1} + I_{CO2}$.

For a specific condition of $I_g = 0$, we have $I = I_A = I_K$, and Eq. 15 reduces to

$$I = M(V_2)[\alpha_1(I)I + \alpha_2(I)I + I_0]. \tag{16}$$

When $I \gg I_0$, it further reduces to the familiar form

$$M(V_2) = \frac{1}{\alpha_1 + \alpha_2}. \tag{17}$$

The multiplication M can generally be related empirically to the junction breakdown voltage V_B as

$$M(V_2) = \frac{1}{1-(V_2/V_B)^n} \tag{18}$$

(see Section 5.2.3) and n is a constant. The forward breakover voltage can now be obtained from Eqs. 17 and 18 (and $V_{AK} \approx V_2$);

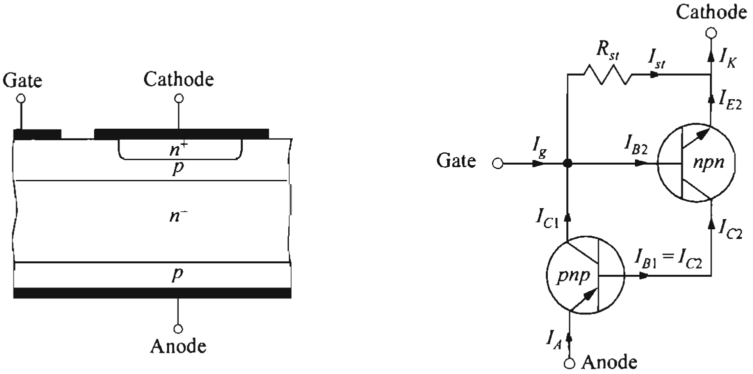
$$V_{BF} = V_B(1-\alpha_1-\alpha_2)^{1/n} \qquad (\alpha_1+\alpha_2) < 1\,. \tag{19}$$

Comparison with the reverse breakdown voltage [$V_{BR} = V_B(1-\alpha_1)^{1/n}$] shows that V_{BF} is always less than V_{BR}. For small values of $(\alpha_1+\alpha_2)$, V_{BF} will be essentially the same as the reverse breakdown voltage shown in Fig. 3. For values of $(\alpha_1+\alpha_2)$ close to 1, the breakover voltages can be substantially less than V_{BR}.

Cathode Short. In modern Shockley-diode and thyristor designs, cathode shorts are often used to improve device performance.[7,8] A schematic diagram of a thyristor with a cathode short is shown in Fig. 10a where the cathode is shorted to the *p*2-region. A two-transistor equivalent circuit is shown in Fig. 10b, where the total cathode current I_K is now the sum of the emitter current I_{E2} and the shunt current I_{st}. The shunt resistance is due to the contact resistance on the *p*-region and the bulk resistance of the *p*-region itself, and it depends on the geometry of the structure. The function of the shunt is to degrade the current gain of the *n-p-n* transistor such that an effective lower value of α_2' should be used in Eq. 19 in place of α_2, to give a larger breakover voltage. The effective current gain with shunt can be shown to be degraded from the original value by;

$$\alpha_2' = \frac{I_{C2}-I_{CO2}}{I_K} = \frac{I_{C2}-I_{CO2}}{I_{E2}+I_{st}} = \frac{\alpha_2}{1+(I_{st}/I_{E2})}\,. \tag{20}$$

Due to the nonlinear dependence of I_{E2} on the base-emitter bias (gate bias), α_2' can be varied from a small value to the original α_2 value. In the extreme case of $\alpha_2' = 0$, the forward breakover voltage can be made as large as the reverse-blocking voltage



(a) **(b)**

Fig. 10 (a) thyristor with cathode short. (b) Circuit representation of 2-transistor analog. Shunt current I_{st} will flow through the cathode short.

as given by Eq. 3. When the thyristor needs to be turned on, the gate bias now has an additional effect on increasing α_2' towards the condition of $(\alpha_1 + \alpha_2') = 1$.

11.2.3 Turn-On Mechanisms

$(\alpha_1 + \alpha_2)$ Criterion. We now go back to the forward-blocking current expression of Eq. 11. As the terminal voltage V_{AK} increases, the total current passing through both *p-n-p* and *n-p-n* transistors will also increase. The higher current will cause α_1 and α_2 to go up (see Fig. 8 on p. 556). Higher current gain will induce even higher current. Because of the regenerative nature of these processes, the device is eventually switched to its *on*-state. This can be seen from Eq. 11 that for the condition of $(\alpha_1 + \alpha_2) = 1$, the anode current will be infinite, i.e., an unstable state where switching occurs.

An increase of α_2 can also be induced by injecting I_g, which is the base current for the *n-p-n* transistor. This explains the lowering of switching voltage as a function of I_g shown in Fig. 2. In the extreme case, for a fixed bias V_{AK}, the gate can be used as a control to turn on and turn off the thyristor.

$(\tilde{\alpha}_1 + \tilde{\alpha}_2)$ Criterion. As discussed above, since the current gain α is a function of the current, it is a variable and so there is a smaller-signal value associated with it. We shall now show that the switching can begin when the sum of the small-signal $\tilde{\alpha}$ reaches unity, which often occurs before that of the dc value.[17] Let us consider the situation that results when the gate current I_g is increased by a small amount ΔI_g. This change will perturb both I_A and I_K, but their difference in changes should be exactly equal to ΔI_g;

$$\Delta I_K - \Delta I_A = \Delta I_g . \tag{21}$$

The small-signal $\tilde{\alpha}$ are defined as

$$\tilde{\alpha}_1 \equiv \frac{dI_{C1}}{dI_A} = \lim_{\Delta I_A \to 0} \frac{\Delta I_{C1}}{\Delta I_A} \tag{22a}$$

$$\tilde{\alpha}_2 \equiv \frac{dI_{C2}}{dI_K} = \lim_{\Delta I_K \to 0} \frac{\Delta I_{C2}}{\Delta I_K} . \tag{22b}$$

The hole current collected by J2 will be $\tilde{\alpha}_1 \Delta I_A$ and the electron current will be $\tilde{\alpha}_2 \Delta I_K$. Equating the change in anode current to the change in current across J2, we obtain

$$\Delta I_A = \tilde{\alpha}_1 \Delta I_A + \tilde{\alpha}_2 \Delta I_K . \tag{23}$$

Substituting Eq. 23 into Eq. 21 yields

$$\frac{\Delta I_A}{\Delta I_g} = \frac{\tilde{\alpha}_2}{1 - (\tilde{\alpha}_1 + \tilde{\alpha}_2)} . \tag{24}$$

When $(\tilde{\alpha}_1 + \tilde{\alpha}_2)$ becomes unity, the device is unstable since from Eq. 24 a small increase in I_g will cause an infinite increase in I_A. Although gate current was used in the analysis, the same effect can be obtained with a slight increase in temperature or voltage.

In the following, we derive the small-signal $\tilde{\alpha}$ and show that it can be larger than the dc α. In this scenario, the criterion of $(\tilde{\alpha}_1 + \tilde{\alpha}_2) = 1$ will occur first. The dc common-base current gain of a transistor is given by

$$\alpha = \alpha_T \gamma \tag{25}$$

where α_T is the transport factor defined as the ratio of the injected current reaching the collector junction to the injected current at the emitter, and γ is the injection efficiency defined as the ratio of the injected minority-carrier current to the total emitter current. By differentiating Eq. 7 with respect to the emitter current, we obtain the small-signal $\tilde{\alpha}$:

$$\tilde{\alpha} \equiv \frac{dI_C}{dI_E} = \alpha + I_E \frac{d\alpha}{dI_E}. \tag{26}$$

Substituting Eq. 25 into Eq. 26 yields

$$\tilde{\alpha} = \gamma \left(\alpha_T + I_E \frac{d\alpha_T}{dI_E} \right) + \alpha_T I_E \frac{d\gamma}{dI_E}. \tag{27}$$

The simplest approximations for α_T and γ are given by (see Chapter 5)

$$\alpha_T = \frac{1}{\cosh(W/L_p)} \approx 1 - \frac{W^2}{2L_p^2}, \tag{28}$$

$$\gamma \approx \frac{1}{1 + (N_B W / N_E W_E)}, \tag{29}$$

where W is the neutral base width (Fig. 3a), L_p is the diffusion length of minority carriers in the base, N_B and N_E are the base and emitter concentrations respectively, and W_E is the emitter length. To obtain small values of α for large V_{BF}, one must use large values of W/L_p and N_B/N_E.

To investigate the dependence of dc α and small-signal $\tilde{\alpha}$ on current, we must use a more-detailed calculation, considering both diffusion and drift current components. We examine the *p-n-p* transistor portion of the thyristor. The hole currents in the base can be calculated from the equation

$$I_p(x) = qA_s \left(p_n \mu_p \mathscr{E} - D_p \frac{dp_n}{dx} \right) \tag{30}$$

where A_s is the area of the junction. The continuity equation for the same region is given by

$$\frac{\partial p_n}{\partial t} = D_p \frac{\partial^2 p_n}{\partial x^2} - \frac{p_n - p_{no}}{\tau_p} - \mu_p \mathscr{E} \frac{\partial p_n}{\partial x}. \tag{31}$$

And the boundary conditions are $p_n(x = \mathrm{J1}) = p_{no}\exp(\beta V_1)$ where $\beta \equiv q/kT$, and $p_n(x = \mathrm{J2}) = 0$. The steady-state solution of Eq. 31 subject to these boundary conditions is[18]

$$p_n(x) = p_{no}\{\exp(\beta V_1)\exp[(C_1 + C_2)x] - [\exp(\beta V_1)\exp(C_2 W) + \exp(-C_1 W)]\exp(C_1 x)\operatorname{csch}(C_2 W)\sinh(C_2 x)\} \quad (32)$$

where

$$C_1 \equiv \frac{\mu_p \mathscr{E}}{2D_p}, \quad (33)$$

$$C_2 \equiv \sqrt{\left(\frac{\mu_p \mathscr{E}}{2D_p}\right)^2 + D_p \tau_p}. \quad (34)$$

From Eqs. 30–32 we obtain for the transport factor

$$\alpha_T \equiv \frac{I_p(x = \mathrm{J2})}{I_p(x = \mathrm{J1})} = \frac{C_2 \exp(C_1 W)}{C_1 \sinh(C_2 W) + C_2 \cosh(C_2 W)}. \quad (35)$$

The injection efficiency is given by

$$\gamma \equiv \frac{I_{pE}}{I_{pE} + I_{nE} + I_r} \approx \frac{I_{pE}}{I_{pE} + I_r} = \frac{I_{po}\exp(\beta V_1)}{I_{po}\exp(\beta V_1) + I_R \exp(\beta V_1/m)} \quad (36)$$

where I_{pE} and I_{nE} are the injected hole and electron currents from the emitter respectively, I_r is the space-charge recombination current given by $I_R\exp(\beta V_1/m)$, where I_R and m are constants (generally $1 < m < 2$), and

$$I_{po} = qD_p A_s p_{no}[C_1 + C_2 \coth(C_2 W)]. \quad (37)$$

For the doping profile of Fig. 1b, $p_{po}(p1) \gg n_{no}(n1)$, the current I_{nE} can be neglected in Eq. 36.

We can now calculate α_1 from Eqs. 35 and 36 as a function of the emitter current. In addition, we can combine Eqs. 27, 35, and 36 to give the small-signal $\tilde{\alpha}$. The results are shown in Fig. 11 for the doping profile similar to that shown in Fig. 1b and for some typical parameters of silicon.[18] Note that for the current range shown, the small-signal $\tilde{\alpha}$ is always greater than the dc α. The ratio of the neutral base width to diffusion length W/L_p is an important device parameter in determining the variation of gain with current. For small values of W/L_p, the transport factor α_T is independent of current, and the gain varies with current only through the injection efficiency. This condition applies to the narrow base-width section of the device (*n-p-n* section). For larger values of W/L_p, both transport factor and injection efficiency are functions of current (*p-n-p* section). Thus, the value of the gain can, in principle, be tailored to the desired range by choosing the proper diffusion length and doping profile.

***dV/dt* Triggering.** Under transient conditions, a forward-blocking thyristor can switch to its *on*-state at voltages well below the breakover voltage. This undesirable effect, which can turn on a thyristor unintentionally under transients, is called *dV/dt* triggering (V is the terminal voltage V_{AK}). The *dV/dt* effect is due to the rapidly varying anode-cathode voltage giving rise to a displacement current across the junction J2, given by $C_2 dV_{AK}/dt$, where C_2 is the depletion capacitance of J2. This displacement current plays the role similar to that of the gate current. It increases the

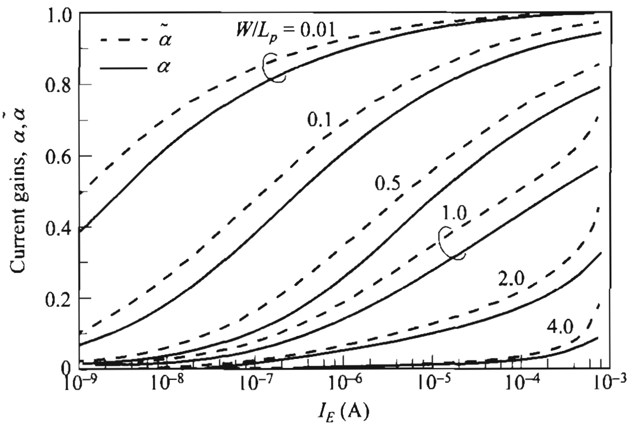


Fig. 11 Small-signal $\tilde{\alpha}$ and dc α as functions of emitter current, for different base width/diffusion length ratio W/L_p. Parameters used: $A_s = 0.16$ mm^2, $L_p = 25.5$ μm. (After Ref. 18.)

small-signal current gains and in turn, can cause $(\tilde{\alpha}_1 + \tilde{\alpha}_2)$ to approach 1; then switching occurs. In power thyristors, *dV/dt* ratings must be high so that false triggering can be avoided.

The origin of the *dV/dt* triggering can be understood as follows. Since in the forward-blocking state, most of V_{AK} will be across the junction J2, a change in V_{AK} will cause a change in the depletion width of J2. To respond to this change, majority carriers on both sides of J2 will flow, resulting in the displacement current. That means electrons in the *n*1-region and holes in the *p*2-region will affect the emitter junctions J1 and J3 respectively. These currents will increase the small-signal current gains $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$.

To improve the *dV/dt* rating, one can reverse bias the gate-cathode terminals so that the displacement current is drawn to the gate from the *p*2-region and will not affect the current gain of the *n-p-n* transistor. The lifetimes in the *n*1- and *p*2-regions can also be deteriorated to reduce the α at any current levels; but this approach will degrade the forward-conduction mode.

An effective method to improve *dV/dt* rating is to use cathode shorts,[19] as shown in Fig. 10. The displacement current (holes) into the J3 junction will be bypassed through the shorts, so α_2 of the *n-p-n* transistor is not affected by this displacement current. The cathode shorts can substantially improve the *dV/dt* capability. Typically a 20-V/μS rating is obtainable in thyristors without cathode shorts. For shorted devices, the *dV/dt* rating can be increased by a factor of 10 to more than 100.

11.2.4 Forward Conduction

The switching between the forward-blocking *off*-state and the forward-conducting *on*-state is depicted in Fig. 12. In equilibrium, there is at each junction a depletion region with a built-in potential that is determined by the impurity doping profile. When a positive voltage is applied to the anode, junction J2 will tend to become

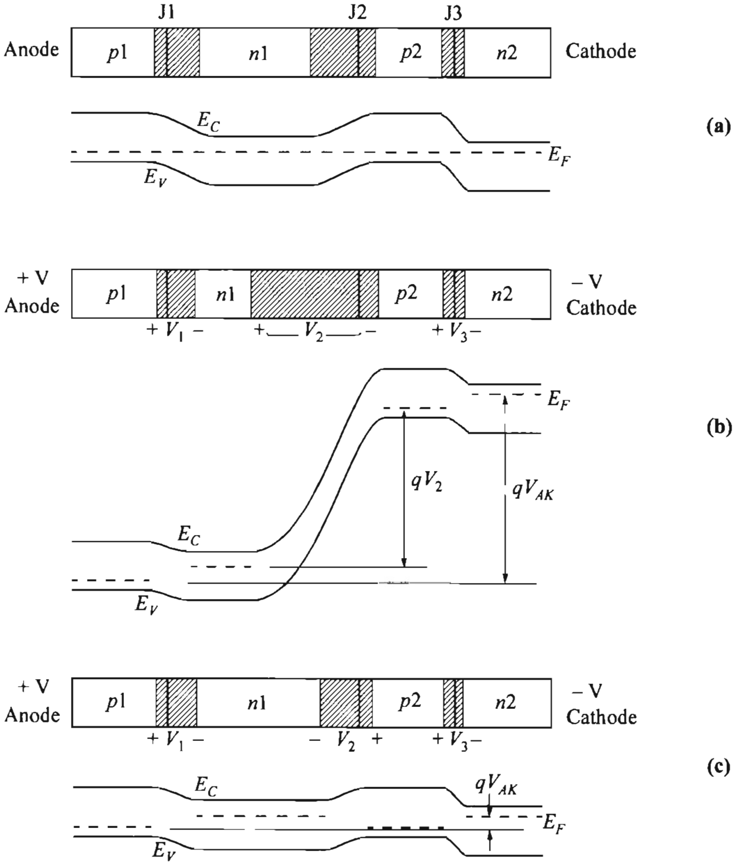


Fig. 12 Energy-band diagrams for forward regions. (a) Equilibrium. (b) Forward *off*-state, where most of the voltage drops across the center junction J2. (c) Forward *on*-state, where all three junctions are forward biased. Note the collapse and reversal of polarity on V_2.

reverse biased, while J1 and J3 will be forward biased. The anode-to-cathode voltage drop is equal to the algebraic sum of the junction drops:

$$V_{AK} = V_1 + V_2 + V_3 . \tag{38}$$

Upon switching, the current through the device must be limited by an external load resistance; otherwise, the device would destroy itself if the supply voltage were sufficiently high. In this *on*-state, J2 is changed from being reverse biased to forward biased, as shown in Fig. 12c, and the net voltage drop V_{AK} is given by $(V_1 - |V_2| + V_3)$ which is approximately equal to the voltage drop across one forward-biased *p-n* junction plus a saturated bipolar transistor.

It is worthwhile to point out that if the polarities of the anode and cathode are reversed, junctions J1 and J3 are reverse biased while J2 is forward biased. Under this condition there is no switching action, since only the center junction acts as an emitter, and no regenerative process can take place. So there is no switching in the reverse V_{AK} polarity.

When the thyristor is in the *on*-state, all three junctions are forward biased. Holes are injected from the *p*1-region and electrons from the *n*2-region. These carriers flood the *n*1- and *p*2-regions which are relatively lightly doped. Therefore, the device behaves like a p^+-i-n^+ (*p*1-*i*-*n*2) diode.

For a p^+-i-n^+ diode with *i*-region width of W_i (W_i is now the sum of the *n*1- and *p*2-regions), the forward current density is accounted for by the rate at which holes and electrons recombine within the W_i region. The current density is thus given by

$$J = q\int_0^{W_i} R dx \tag{39}$$

where R is the recombination rate that can be expressed as[20]

$$R = A_r(n^2p + p^2n) + \frac{np - n_i^2}{\tau_{po}(n + n_i) + \tau_{no}(p + n_i)}. \tag{40}$$

The first term is due to Auger processes and the Auger coefficient A_r is found to be $1\text{–}2\times10^{-31}$ cm^6/s for silicon; the second term is due to midgap recombination traps, and τ_{po} and τ_{no} are the hole and electron lifetimes, respectively. Under high-level injection, $n = p \gg n_i$, Eq. 40 reduces to

$$R = n\left(2A_rn^2 + \frac{1}{\tau_{po} + \tau_{no}}\right). \tag{41}$$

If the carrier concentration is approximately constant throughout the W_i-region, the current density from Eqs. 39 and 41 can be written as

$$J = \frac{qnW_i}{\tau_{\text{eff}}}, \tag{42}$$

where the effective lifetime is

$$\tau_{\text{eff}} = \frac{n}{R} = \left(2A_rn^2 + \frac{1}{\tau_{po} + \tau_{no}}\right)^{-1}. \tag{43}$$

We now examine the voltage dependence to give the *I-V* characteristics. To gain some physical insight, we first look at the internal voltage drop V_i across the W_i-region. Treating the problem as a drift process, we can interpret the current as

$$J = q(\mu_n + \mu_p)n\overline{\mathscr{E}} \tag{44}$$

where $\overline{\mathscr{E}}$ is the average electric field. Since $V_i = W_i\overline{\mathscr{E}}$, from Eqs. 42 and 44, we can obtain the internal voltage drop as

$$V_i = \frac{W_i^2}{(\mu_n + \mu_p)\tau_{\text{eff}}}. \tag{45}$$

Because V_i is inversely proportional to the effective lifetime, longer τ_{eff} is desirable. Calculated values of τ_{eff} are shown in Fig. 13 as a function of injected concentration,

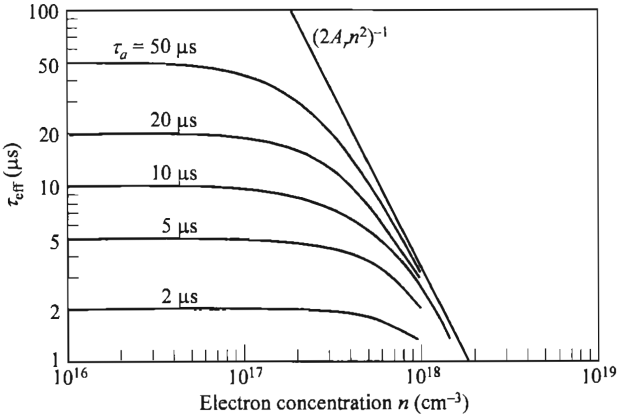


Fig. 13 Effective lifetime under high-injection condition. τ_a is the ambipolar lifetime and A_r is the Auger coefficient (= 1.45×10^{-31} cm^6/s). (After Ref. 8.)

for different values of ambipolar lifetime $\tau_a = (\tau_{po} + \tau_{no})$. At low carrier concentrations, the effective lifetime is the same as the ambipolar lifetime; however, at carrier concentrations above 10^{17} cm^{-3}, the effective lifetime falls rapidly as n^{-2} due to Auger processes. Also, at high carrier concentrations, the additional effect of carrier-carrier scattering also starts due to strong interaction between mobile carriers. This effect is interpreted through a parameter called the ambipolar diffusion coefficient, given by

$$D_a = \frac{n+p}{n/D_p + p/D_n}. \tag{46}$$

Equation 45 can be rewritten as

$$V_i = \frac{2kTbW_i^2}{q(1+b)^2 D_a \tau_{\text{eff}}} \tag{47}$$

where b is the ratio $\mu_n/\mu_p = D_n/D_p$. At low n and p concentrations,

$$D_a = \frac{2D_n D_p}{D_n + D_p}, \tag{48}$$

and is independent of carrier concentration. The effect of carrier-carrier scattering is included in D_a whose dependence on excess carrier concentration is shown in Fig. 14. From this discussion, we see that V_i increases with current (or n) indirectly through both τ_{eff} and D_a.

The total terminal voltage drop should also include the end regions and their injection efficiencies. With these effects, the terminal I-V relationship is given as[8]

$$J = \frac{4qn_i D_a F_L}{W_i} \exp\left(\frac{qV_{AK}}{2kT}\right). \tag{49}$$

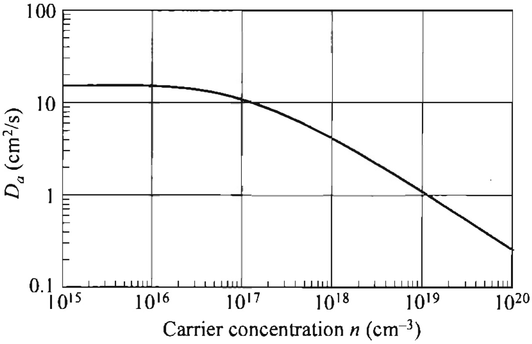


Fig. 14 Ambipolar diffusion coefficient as a function of injected carrier concentration. (After Ref. 8.)

The factor of 2 in the exponential term is characteristic of a recombination process. The value of F_L is a function of W_i/L_a where L_a is the ambipolar diffusion length $L_a = (D_a\tau_a)^{1/2}$, and its dependence is shown in Fig. 15.

A similar equation can be obtained quickly from a simple recombination/generation consideration, which can provide some physical insight. The recombination current within a depletion region is given by

$$J_{re} = \frac{qW_in_i}{2\tau}\exp\left(\frac{qV}{2kT}\right) \tag{50}$$

(see Section 2.3.2). Assuming that W_i is comparable to the ambipolar diffusion length, $W_i \approx \sqrt{D_a\tau}$. Substitution of τ into Eq. 50 gives

$$J_{re} \approx \frac{qn_iD_a}{2W_i}\exp\left(\frac{qV}{2kT}\right). \tag{51}$$

This result is not very different from Eq. 49.

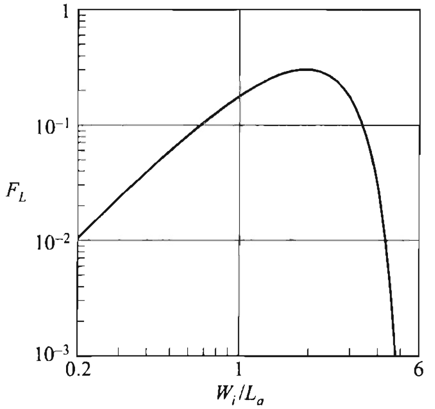


Fig. 15 F_L as afunction of W_i/L_a. (After Ref. 8.)

Numerical analysis for the forward conduction has been done incorporating various physical mechanisms. A series of calculated *I-V* curves for a 2.5-kV thyristor are shown[20] in Fig. 16 for a heat-sink temperature of 400 K. The inscription of each curve indicates the physical mechanisms that were removed. For example, "carrier-carrier removed" means the removal of carrier-carrier scattering in the numerical analysis. The l-kA/cm^2 level is associated with maximum surge operation while 100-A/cm^2 level is associated with maximum steady-state operation. As can be seen, carrier-carrier scattering and Auger recombination are important limiting mechanisms at both of these operation levels. The bandgap narrowing has virtually no effect until the current density is above 1 kA/cm^2. The midgap trap recombination becomes the limiting factor at levels below 100 A/cm^2, and is also important at the surge level. The junction-temperature effect becomes important when the current density is larger than 500 A/cm^2. The bottom curve is the *nominal* case incorporating all the mechanisms described above. Also shown are the experimental results, which are in excellent agreement with the nominal case.

***dI/dt* Limitation.** Initially, in the thyristor turn-on process, only a small area of the cathode region near the gate contact begins to conduct.[8] This highly conducting region supplies the necessary forward current to turn on adjacent regions until the conduction process spreads over the entire cross section of the cathode. The spreading of the conduction process is limited and characterized by a spreading velocity v_{sp}. If the anode current is increased too rapidly during the turn-on process, a large current density results at the edge of cathode which is turned on first by the gate

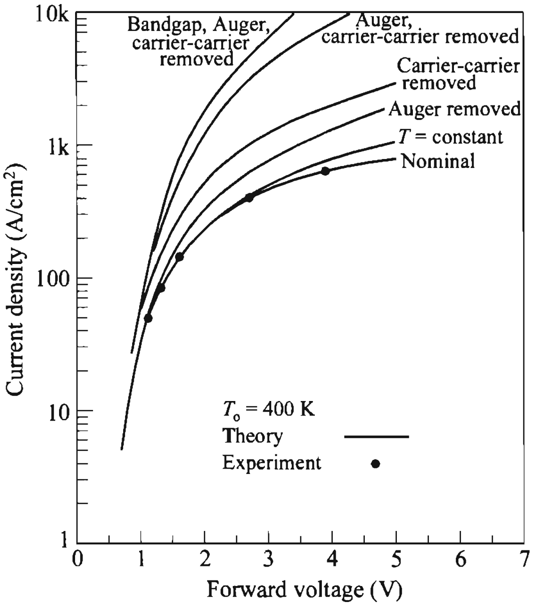


Fig. 16 Theoretical curves illustrating the relative importance of various physical mechanisms including heat flow, on the *I-V* characteristics of a 2.5-kV thyristor. Thermal conductivity = 50 W/cm^2-K. Also shown are the measured results. (After Ref. 20.)

injection. The high current density results in a hot spot and permanent damage can result.

The problem becomes a race between the rate of increase of the total current versus the rate by which the effective cathode area is expanding (v_{sp}). The local temperature of the hot spot can be estimated by the power density, given by

$$\Delta T = \frac{\text{power}}{\text{effective cathode area}} \propto \frac{dI_A/dt}{v_{sp}^k} . \tag{52}$$

The constant k depends on the geometry of the gate and the cathode. It is ≈ 1 for linear cathode strips, and ≈ 2 for a circular, concentric gate and cathode layout. The spreading velocity v_{sp} is typically under 10^4 cm/s. It has been found to increase with triggering gate current I_g, and decrease with the total width W_i. The latter constraint on W_i imposes a compromise between breakdown voltage and dI_A/dt rating. Equation 52 shows that for a given device, the temperature rise is proportional to dI_A/dt. The allowable dI_A/dt is, thus, an important rating.

For a given device, one can minimize the problem by overdriving the device with a high triggering gate current. Another obvious circuit approach is to add inductance in series to the anode/cathode terminal to limit the fast transience being fed to the device. The following discusses a couple of device designs for a better dI/dt capability.

A number of interdigitated (between gate and cathode) designs have been developed so that no part of the cathode area is greater than a certain maximum allowed distance from the gate electrode. A simple structure consists of thin and long gate and cathode strips. A more-complex design is a involute pattern which consists of spiral gate and cathode strips with constant width and spacing between them.[21]

Another method to enlarge the initial turn-on area is the use of an amplifying gate (Fig. 17).[22] When a small triggering current is applied to the central gate, the amplifying gate structure, which serves as a pilot parasitic SCR, will turn on rapidly because of its small lateral dimensions. The pilot current is much larger than the original triggering gate current and it provides a much-higher driving gate current to the main device. As pointed out before, the larger the driving gate current is, the larger is the initial turn-on area for the main thyristor. This design effectively employs a small parasitic SCR to amplify the gate current internally to improve the dI/dt rating.

11.2.5 Static *I-V* Curves

After discussing each mode of operation in different bias regimes, we now examine the complete set of I-V characteristics. We start with a simpler case of two-terminal Shockley diode. From the general equations we can develop a method to generate the I-V characteristics.[23] Since $I_g = 0$ and $I_A = I_K = I$ in a Shockley diode, Eqs. 15 and 18 give

$$\frac{1}{M(V_2)} = \alpha_1(I) + \alpha_2(I) + \frac{I_0}{I} = 1 - \left(\frac{V_2}{V_B}\right)^n . \tag{53}$$

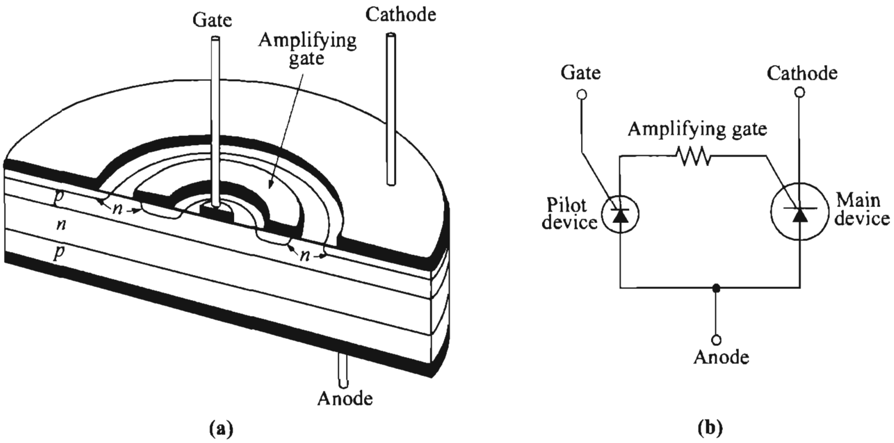


Fig. 17 (a) SCR with amplifying gate for improved dI/dt. (b) Its equivalent circuit. (After Ref. 22.)

We shall assume that I_o is some known constant, and α_1 and α_2 are known functions of current similar to that shown in Fig. 11. So for a given current I, Eq. 53 gives the corresponding value in V_2/V_B. The qualitative result is shown in Fig. 18a.

Note that from the figure the switching point (I_s, V_{BF}) occurs at the location where Eq. 53 reaches its minimum (or maximum V_2/V_B) at which point the magnitude of I_s can be calculated. After switching, the holding point is defined as the low-voltage, high-current point at which $dV/dI = 0$. As mentioned before, all three junctions will be under forward bias, and this analysis does not enable us to find this holding point since Eq. 53 is not applicable for a forward-biased J2. However, we can still estimate the holding voltage V_h as follows. When the device is turned on, the net terminal

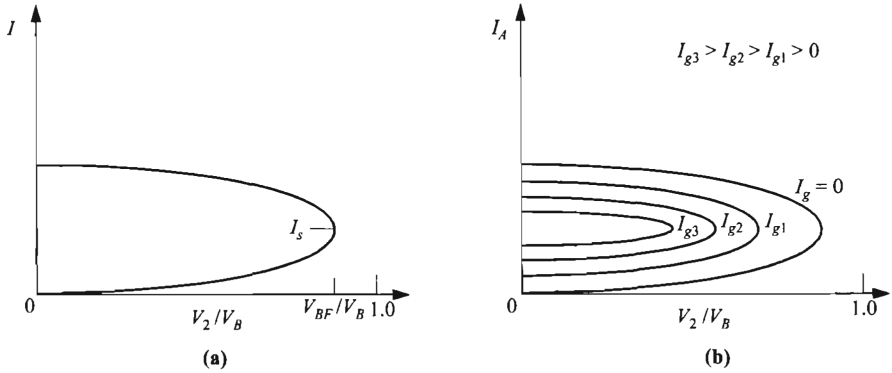


Fig. 18 Solution of I-V characteristics: (a) Two-terminal Shockley diode. (b) Three-terminal thyristor.

voltage V_{AK} is the sum of three forward-biased *p-n* junctions, with the middle J2 being negative in the algebraic sum, as shown in Fig. 12c. Alternatively, the terminal voltage is one forward junction plus the V_{CE} of a bipolar transistor in saturation. These values are approximately 0.7 V and 0.2 V respectively, leading to a holding voltage V_h of ≈ 0.9 V. Beyond this point the device is in forward conduction as discussed in Section 11.2.4.

For the thyristor with the third-terminal gate electrode, Eq. 53 becomes

$$\frac{1}{M(V_2)} = \alpha_1(I_A) + \alpha_2(I_A + I_g) + \frac{\alpha_2(I_A + I_g)}{I_A} I_g + \frac{I_0}{I_A} = 1 - \left(\frac{V_2}{V_B}\right)^n . \tag{54}$$

In getting to Eq. 54 the current I_K is replaced by $I_A + I_g$, and the term $\alpha_2(I_A + I_g)I_g/I_A$ is included. $\alpha_2(I_A + I_g)$ for each value of I_g is first reevaluated. Following the previous procedure, it generates a set of *I-V* curves shown in Fig. 18b. We note that as I_g increases, the switching voltage decreases. This gives rise to the general gate turn-on properties of the thyristor.

The complete terminal *I-V* characteristics for the gate-triggered thyristor are shown in Fig. 2 for a family of different gate currents. In the forward-blocking state, the curves are similar to those shown in Fig. 18b except for a change of coordinates.

11.2.6 Turn-On and Turn-Off Times

Turn-On Time. To switch a thyristor from *off* to *on* requires that the current be raised to a level high enough to satisfy the condition $(\alpha_1 + \alpha_2)$ [or $(\tilde{\alpha}_1 + \tilde{\alpha}_2)$] = 1. A number of methods can be used to trigger thyristors from the *off*-state to the *on*-state. Voltage triggering is the only method of switching a two-terminal Shockley diode. Voltage triggering can be accomplished in two ways: by slowly raising the forward voltage until the breakover voltage is reached, or by applying the anode voltage rapidly, referred to as *dV/dt* triggering, considered in Section 11.2.3.

Gate-current triggering is the most-important method of switching a three-terminal thyristor. When a triggering current (e.g., gate current) is applied, the anode current through a thyristor does not respond immediately. The anode current can be characterized by two transition times as shown in Fig. 19a. The first component is the delay time t_d which is associated with the intrinsic speed of the two bipolar transistors. This delay is the sum of the base transit times

$$t_d = t_1 + t_2 \tag{55}$$

where $t_1 = W_{n1}^2/2D_p$, $t_2 = W_{p2}^2/2D_n$, W_{n1} and W_{p2} are the layer widths of the *n*1- and *p*2-regions, respectively, and D_p and D_n are the hole and electron diffusion coefficients respectively.

The second component, the rise time t_r, is related to the build-up of the stored charges Q_1 and Q_2 within the base regions of the *p-n-p* and *n-p-n* transistors, once they are turned on. The collector currents in the transistors are related to these charges by $I_{C1} \approx Q_1/t_1$ and $I_{C2} \approx Q_2/t_2$. Because of the regenerative nature of a thyristor, the rise time is approximately the geometric mean value of the diffusion times in the *n*1- and *p*2-regions, or

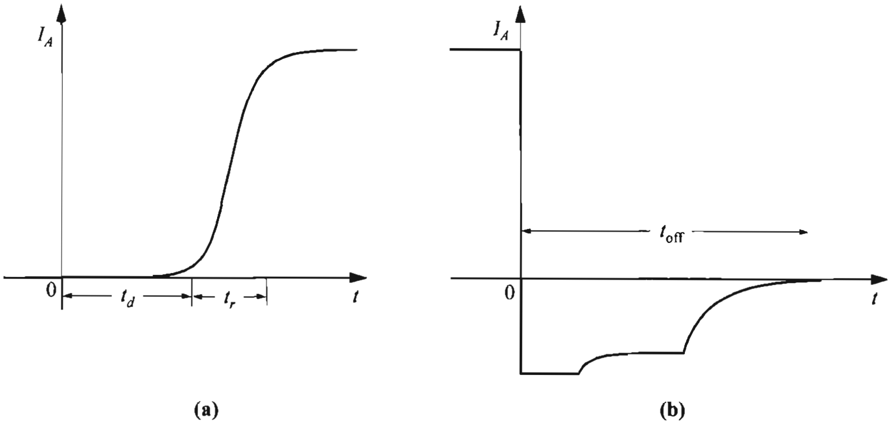


Fig. 19 (a) Turn-on characteristics triggered by a gate-current step I_g ($t = 0$) applied to a thyristor. (b) Turn-off characteristics where the voltage V_{AK} suddenly changes polarity.

$$t_r = \sqrt{t_1 t_2}. \tag{56}$$

This result can be derived from Fig. 8b with the help of the charge-control approach. Under the ideal condition that $dQ_1/dt = I_{B1} = I_{C2}$ and $dQ_2/dt = I_{B2} = I_g + I_{C1}$, we obtain the following equation:

$$\frac{d^2Q_1}{dt^2} - \frac{Q_1}{t_1 t_2} = \frac{I_g}{t_2}. \tag{57}$$

The solution of Eq. 57 is of the form $Q_1 \propto \exp(-t/t_r)$, with the time constant t_r given by Eq. 56. To reduce the total turn-on time, one must employ devices with narrow $n1$- and $p2$-layer widths. This requirement, however, is in contrast to that for high breakdown voltage, and is the reason that high-power, high-voltage thyristors usually have longer turn-on times.

Turn-Off Time. When a thyristor is in the *on*-state, all three junctions are forward biased. Consequently in the device, excess minority and majority carriers exist and increase with forward current. To switch back to the blocking state, these excess carriers must be swept out by an electric field or must decay by recombination.[24,25] A typical turn-off current waveform is shown in Fig. 19b, where the terminal voltage V_{AK} is suddenly changed to the opposite polarity. While exact analysis of the current waveform in Fig. 19b is complicated, one can get a simple estimate of the turn-off time by the following.[7] The major time delay can be considered due to the recombination time in the $n1$-layer. The base-charge recombination is governed by

$$\frac{dQ_1}{dt} + \frac{Q_1}{\tau_p} = 0. \tag{58}$$

Since the hole current through the structure is proportional to this base charge, we can write the solution of Eq. 58 as

$$Q_1(t) = Q_1(0)\exp\left(\frac{-t}{\tau_p}\right)$$
$$= \tau_p\alpha_1 I_F\exp\left(\frac{-t}{\tau_p}\right), \tag{59}$$

where we obtain the initial base charge $Q_1(0)$ being proportional to the forward-conduction current I_F. The anode current is expected to decay according to

$$I_A = I_F\exp\left(-\frac{t}{\tau_p}\right), \tag{60}$$

where τ_p is the minority-carrier lifetime in the *n*1-base. This current must drop below the holding current I_h to permit the device to switch to the forward-blocking state. Thus the turn-off time is given by

$$t_{\text{off}} = \tau_p \ln\left(\frac{I_F}{I_h}\right). \tag{61}$$

To obtain a small turn-off time, we must reduce the lifetime τ_p in the *n*1-layer. This reduction can be achieved by introducing recombination centers, such as gold and platinum, in silicon during the diffusion process, or using electron and gamma-ray irradiation.[26–27] Gold has an acceptor level near the midgap of silicon to serve as an efficient generation-recombination center but the leakage current increases. As a result, the forward breakover voltage decreases with gold doping. The reduction in lifetime will also cause the forward voltage drop in the *on*-state (Eq. 47) to increase. So there is a trade-off between the forward voltage drop and turn-off time for proper optimization for specific applications.

To shorten the turn-off time, a common circuit practice is to apply a reverse bias between the gate and the cathode during the turn-off phase, in addition to reversing the polarity of V_{AK}. This method is called gate-assisted turn-off[28,29] (gate turn-off will be discussed in Section 11.3.1). The improvement comes about because the reverse-biased gate can divert most of the forward recovery current which would otherwise flow through the cathode during the reapplication of the forward anode voltage.

Maximum Operating Frequency. At low switching speeds, the thyristor is generally a more-efficient switch than a bipolar transistor. The thyristor has essentially dominated the field of industrial power control, where the operating frequency is usually 50 or 60 Hz. Recently, the development of circuit application for higher switching speeds has increased. We shall now consider the maximum operating frequency obtainable in thyristors.

The operation frequency obviously is limited at least by the turn-on and turn-off times. In practice, the turn-off time is longer and it should be the dominating factor of the two. However, there are two other effects that limit the maximum operation frequency.[30] The first is false triggering caused by *dV*/*dt* transients. The rate at which the forward voltage can be reapplied to the thyristor *dV*/*dt* after a reverse recovery period is limited by the capacitive displacement current. This displacement current may cause the α_2 of the *n-p-n* transistor to rise sufficiently to switch on the thyristor before the full forward voltage has been applied and before any signal has been applied to the gate. This effect can be substantially reduced by cathode shorts. Secondly, the rate of change of current *dI/dt* in the device during switching is a major factor affecting turn-on and turn-off times of thyristors. The *dI/dt* rate is controlled primarily by external circuitry, and it is necessary to ensure that the *dI/dt* ratings, as discussed in the previous section, are not exceeded, to avoid permanent damage.

With these factors considered, the total forward recovery time t_{fr}, the time passed before a high voltage can be applied again without turning on the device, is the sum of the above three components:

$$t_{fr} = t_{\text{off}} + \frac{I_F}{dI/dt} + \frac{V_{BF}}{dV/dt}. \tag{62}$$

The maximum operating frequency thus is given by

$$f_m \approx \frac{1}{2t_{fr}}. \tag{63}$$

In general, t_{off} increases linearly with τ_p in accordance with Eq. 61, so a short τ_p will improve the frequency limit. Meanwhile, a short τ_p will degrade the forward-blocking voltage. As a result, the power rating usually varies inversely as the frequency capability. Also, to improve f_m, both the *dV/dt* and *dI/dt* ratings would have to be optimized or the complexity of the device structure increased.

11.3 THYRISTOR VARIATIONS

11.3.1 Gate Turn-Off Thyristor

A gate turn-off thyristor (GTO) is a thyristor designed to be turned on with a positive gate current and turned off with a negative gate current, while the anode-cathode voltage V_{AK} remains constant in the forward mode. A conventional thyristor is turned off generally by reducing the anode current to below the holding current, or by reversing the V_{AK} polarity and anode current. A GTO can be used for applications in inverter, pulse generator, chopper, and dc switching circuits. The GTO is often used in preference to transistor in high-speed, high-power applications because of its ability to withstand higher voltage in its *off*-state.

A schematic GTO biasing circuit is shown in Fig. 20a. In a one-dimensional description of the turn-off process, one can consider a GTO having a negative gate current of magnitude I_g^-. Referring to Fig. 8b and neglecting all leakage currents, the base drive required to sustain the *n-p-n* transistor in its *on*-state is equal to $(1 - \alpha_2)I_K$. The actual base current is $(\alpha_1 I_A - I_g^-)$. Therefore, the turn-off condition is

$$(1 - \alpha_2)I_K > \alpha_1 I_A - I_g^-. \tag{64}$$

Since $I_A = I_K + I_g^-$, the required I_g^- is given from Eq. 64:

$$I_g^- > \left(\frac{\alpha_1 + \alpha_2 - 1}{\alpha_2}\right) I_A. \tag{65}$$

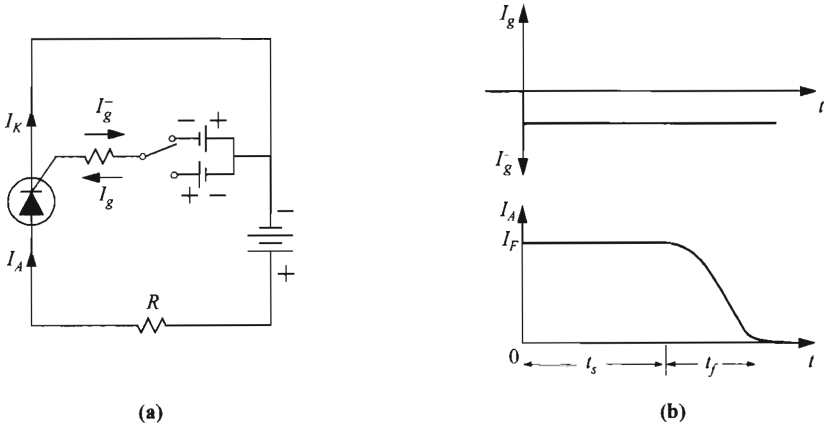


Fig. 20 (a) Biasing circuit diagram for gate turn-off thyristor GTO. (b) Turn-off characteristics of a GTO.

We shall define the ratio I_A/I_g^- for minimum I_g^- as the turn-off gain β_{off}:

$$\beta_{\text{off}} \equiv \frac{I_A}{I_g^-} = \frac{\alpha_2}{\alpha_1 + \alpha_2 - 1} . \tag{66}$$

A high β_{off} means requirement of only a smaller I_g^- to turn off the thyristor. It can be obtained by making α_2 of the *n-p-n* transistor as close to unity as possible, and, at the same time, making α_1 of the *p-n-p* transistor small.

In an actual thyristor, the turn-off is a two-dimensional process. Prior to the applied negative gate current, both transistors are heavily saturated in their *on*-state. The removal of excess stored charges is an important part of the turn-off process. This removal of stored charges results in a storage time delay t_s, followed by a fall time t_f (Fig. 20b), after which the thyristor is in its *off*-state.

As soon as a negative bias is applied to the gate, stored charges will be removed from the $p2$-region by the gate current. This removal is an inverse of the current spreading during the turn-on process. Because of the voltage drop due to the lateral current in $p2$-region, the junction J3 becomes less positively biased as we proceed from the center of the device toward the gate contact (Fig. 21). Eventually, the portion of J3 that is closest to the gate contact becomes reverse biased. At this point, all the forward current will be squeezed into the remaining portion of J3 that is still forward biased. The forward current will be progressively squeezed into a smaller and smaller region until some limiting dimension is reached. At that limit, the remaining excess charge in $p2$-region is removed, and the storage phase is over. The storage time is given by the expression[31]

$$t_s = t_2(\beta_{\text{off}} - 1)\ln\left(\frac{SL_n/W_{p2}^2 + 2L_n^2/W_{p2}^2 - \beta_{\text{off}} + 1}{4L_n^2/W_{p2}^2 - \beta_{\text{off}} + 1}\right) \tag{67}$$

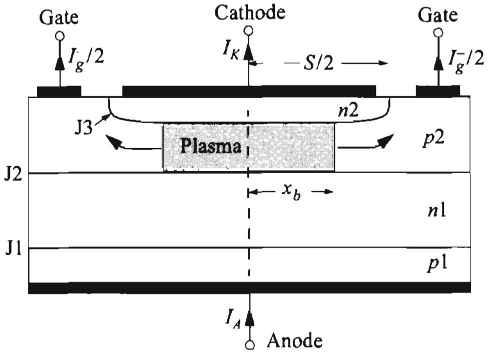


Fig. 21 Plasma storage in the p-base of a gate turn-off thyristor. (After Ref. 31.)

where t_2 is the transit time through the $p2$-base of the n-p-n transistor ($= W_{p2}^2/2D_n$), L_n the electron diffusion length, S the length of the cathode electrode, and W_{p2} is the width of $p2$-region. The storage time will increase with increasing turn-off gain β_{off}, so there is a trade-off between the storage time and the turn-off gain. To reduce the storage time, low β_{off} (corresponding to large gate current) would be preferred.

The fall time in Fig. 20b corresponds to the time required to expand the depletion layer across J2 into the $n1$-region and to remove the hole charges in this region. The total charge per unit area in this $n1$-region is

$$Q \approx qp^* W_D(V_{AK}) \approx J_F t_f \tag{68}$$

where p^* is the average hole concentration in the $n1$-region, W_D the depletion-layer width for a given anode-cathode voltage V_{AK}, and J_F the forward-conduction anode current density. From Eq. 68 the fall time is given by

$$t_f \approx \frac{qp^* W_D(V_{AK})}{J_F} \approx \frac{p^*}{J_F}\sqrt{\frac{2q\varepsilon_s V_{AK}}{N_D}}, \tag{69}$$

where N_D is the doping level in the $n1$-region. The fall time decreases with increasing anode current density and increases with $\sqrt{V_{AK}}$.

Reliable operation of GTO can be obtained when the final area of the squeezed plasma is large enough to prevent excessive current density. This requirement has resulted in the use of interdigitated designs, such as the involute structure.[21] The use of an amplifying gate is also desirable to achieve fast turn-on.

The main difference between the GTO and the previously discussed gate-assisted turn-off thyristor is that the former can be turned off by the application of a negative bias on the gate, while the anode is kept positive with respect to the cathode. On the other hand, the latter requires commutation of the supply voltage to turn it off, and a reverse gate bias helps to reduce the turn-off time.

11.3.2 Diac and Triac

The diac (diode ac switch) and triac (triode ac switch) are bidirectional thyristors.[32,33] They have *on*- and *off*-states for both positive and negative terminal voltages and are therefore useful in ac applications.

The two diac structures are the ac trigger diode and the bidirectional *p-n-p-n* diode switch. The former is simply a three-layer device similar in construction to a bipolar transistor (Fig. 22a), except that doping concentrations at the two junctions are approximately the same and no contact is made to the middle base region. The equal doping levels result in a symmetrical, bidirectional characteristic as shown in Fig. 22c. When a voltage of either polarity is applied to a diac, one junction is forward biased and the other is reverse biased. The current is limited by the leakage current of the reverse-biased junction. When the applied voltage is sufficiently high, breakdown occurs at $V_{BCBO}(1 - \alpha)^{1/n}$, where V_{BCBO} is the avalanche breakdown voltage of the *p-n* junction, α the common-base current gain, and n is a constant. This expression is the same for the breakdown voltage of an open base *n-p-n* bipolar transistor (refer to

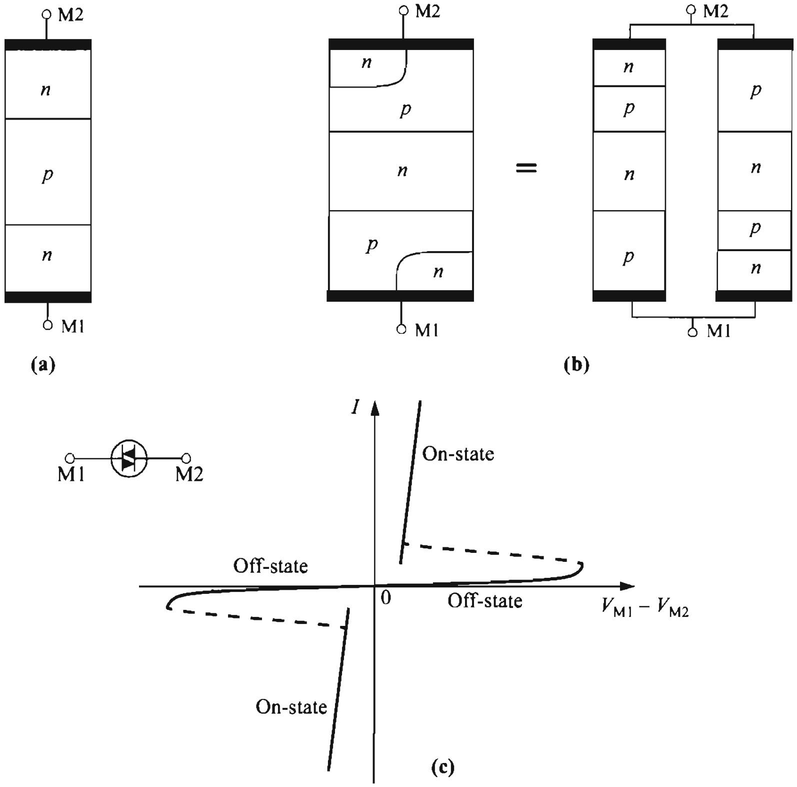


Fig. 22 Structures of diac; (a) ac trigger diode (*n-p-n*) and (b) *n-p-n-p-n* structure equivalent to two Shockley diodes connected in antiparallel. (c) Diac *I-V* characteristics and the device symbol (inset).

Section 5.2.3). As the current increases after breakdown, α will increase, causing a reduction of the terminal voltage. This reduction gives rise to the negative differential resistance.

The bidirectional *p-n-p-n* diode switch behaves like two conventional Shockley diodes connected in antiparallel to accommodate voltage signals of both polarities (Fig. 22b). Using the shorted-cathode principle, we can integrate this arrangement into a single two-terminal diac as shown. The symmetry of this structure results in identical performance for either polarity of applied voltage. The symmetrical *I-V* characteristics and the device symbol are shown in Fig. 22c. Similar to the Shockley diode, the diac can be triggered into conduction by exceeding the breakover voltage or by *dV/dt* triggering. Because of its regenerative action, the bidirectional *p-n-p-n* diode switch has a larger negative resistance and smaller forward drop than that of an ac trigger diode.

The triac is a diac with a third-terminal gate contact to control the switching voltages in both M1-M2 voltage polarities (Fig. 23). The triac structure is considerably more complicated than a conventional thyristor. In addition to the *p*1-*n*1-*p*2-*n*2 basic four layers, there exist at the junction gate an *n*3-region and another *n*4-region in contact with M1. Note also that *p*1 is shorted to *n*4, *p*2 to *n*2, and *n*3 to *p*2, by the three electrodes separately. The triac is very useful in light dimming, motor speed control, temperature control, and other ac applications.

The current-voltage characteristics of a triac are shown in Fig. 23b. The device operations under various biasing conditions are explained in Fig. 24. When the main terminal M1 is positive with respect to M2 and a positive voltage is applied to the gate (also with respect to M2), the device behavior is identical to that of a conventional thyristor (Fig. 24a). The junction J4 is partially reverse biased (due to *IR* drop locally)

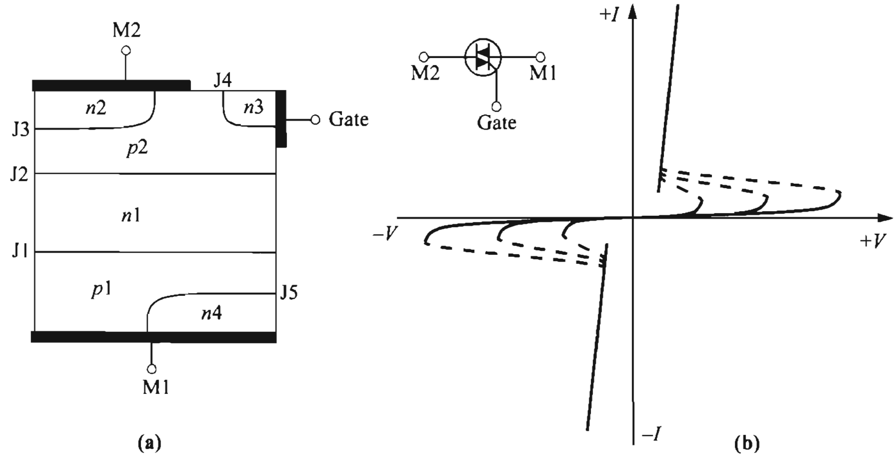


Fig. 23 (a) Cross-section view of a triac, a six-layer structure having five *p-n* junctions (J1–J5) and three electrode shorts. (b) Triac *I-V* characteristics for varying gate biases and the device symbol (inset).

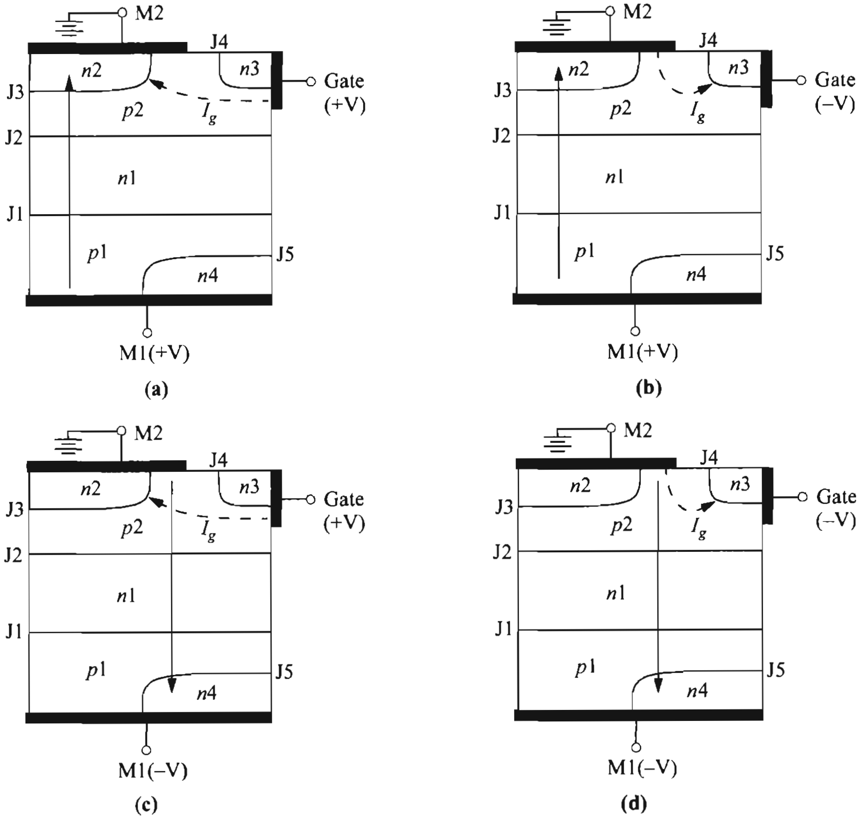

M2
J4
n2
n3
Gate (+V)
J3
p2
I_g
J2
n1
J1
p1
J5
n4
M1(+V)
(a)
M2
J4
n2
n3
Gate (−V)
J3
p2
I_g
J2
n1
J1
p1
J5
n4
M1(+V)
(b)
M2
J4
n2
n3
Gate (+V)
J3
p2
I_g
J2
n1
J1
p1
J5
n4
M1(−V)
(c)
M2
J4
n2
n3
Gate (−V)
J3
p2
I_g
J2
n1
J1
p1
J5
n4
M1(−V)
(d)

Fig. 24 Current-flow patterns in four different triggering modes of a triac. (After Ref. 32.)

and is inactive; the gate current is supplied through the gate short. Since junction J5 is also (partially) reverse biased and inactive, the main current is carried through the left side of *p*1-*n*1-*p*2-*n*2 section.

In Fig. 24b, M1 is positive with respect to M2, but a negative voltage is applied to the gate. The junction J4 between *n*3 and *p*2 is now partially forward biased (due to *IR* drop), and electrons are injected from *n*3 to *p*2. The auxiliary thyristor *p*1-*n*1-*p*2-*n*3 will be turned on by the flow of the lateral base current in *p*2 toward the *n*3-gate because of the gain increase in the transistor *n*3-*p*2-*n*1. Full conduction of this auxiliary thyristor results in the current flowing out of this device and toward the *n*2-region. This current will provide the required gate current and trigger the left-side *p*1-*n*1-*p*2-*n*2 thyristor into conduction.

When M1 is negatively biased with respect to M2, and V_G is positively biased, the junction J3 becomes forward biased between M2 and the shorted gate (Fig. 24c). Electrons are injected from *n*2 to *p*2 and diffuse to *n*1, resulting in an increase of the forward bias of J2. By the regenerative action, eventually full current is carried through the short at M2. The gate junction J4 is reverse biased and is inactive. The full device current is carried through the right-side $p2$-$n1$-$p1$-$n4$ thyristor.

Figure 24d shows the condition for M1 negative with respect to M2 and V_G is also negative. In this condition, the junction J4 is forward biased, and triggering is initiated by injection of electrons from $n3$ to $n1$. This action lowers the potential at $n1$, causing holes to be injected from $p2$ into $n1$. These holes provide the base current drive for the $p2$-$n1$-$p1$ transistor, and the right-side $p2$-$n1$-$p1$-$n4$ thyristor is eventually turned on. Since J3 is reverse biased, the main current is carried from the short at M2 through the $n4$-region.

The triac is a symmetrical triode switch that can control loads supplied with ac power. The equivalence of integrating two thyristors on a single chip results in only half the structure being used at any one time (Fig. 24). Therefore, triac area utilization is poor—about equal to that of two independently connected thyristors. The main advantages of the device are its perfect matching of output characteristics, and the elimination of one package and additional external connections. However, their input characteristics are grossly mismatched. Triacs now have encompassed a wide range of operating voltages (up to 1.6 kV) and currents (over 300 A).

11.3.3 Light-Activated Thyristor

The light-activated thyristor (LASCR), also called the light-activated switch, is a two-terminal (gate contact is optional) four-layer forward-blocking thyristor which can be turned on by exceeding its light-intensity threshold level. The LASCR enables perfect electrical isolation between power and trigger circuits through the use of fiber-optics transmission of the trigger energy. Its applications include photoelectric control such as for street lamps, position monitoring, card readers, and light coupled and triggering circuits.

A simplified LASCR device structure is shown in Fig. 25. The cathode area is exposed homogeneously to a light source through an optical fiber up to a radius r_1,

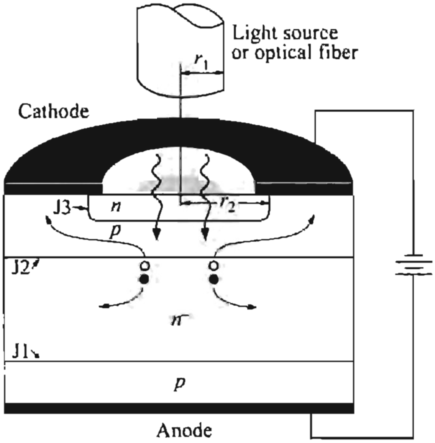


Fig. 25 A light-activated thyristor having a cathode short under illumination.

resulting in a uniform generation of electron-hole pairs within the irradiated area. The LASCR does not require a gate terminal, but the equivalence of the gate current is supplied by an internal photocurrent. Most LASCRs incorporate a cathode short to improve the *dV/dt* tolerance, but it also raises the light power required to trigger the device. Light is mostly absorbed in the wide space-charge region of the central n^--layer where electron-hole pairs are generated. The generated holes are attracted to the cathode short around the *n*-diffusion and produce an *IR* drop to forward bias the cathode emitter junction J3 so electrons can be injected from the *n*-cathode. The generated electrons remain in the n^--layer and supply the base current for the *p-n-p* transistor. Both of these mechanisms help to trigger the thyristor. The light power required for switching is on the order of mW. The advantages of the LASCR include complete electrical isolation from the trigger circuit, and being a more compact and inexpensive integration of a photodetector and an SCR. The turn-on (of the order of 1 μs) is also faster than a regular SCR since the internally generated gate current is more evenly distributed within the device.

At the moment of light exposure the anode current increases abruptly by the amount of photocurrent I_{ph} as shown in Fig. 26. The photocurrent is then amplified in the two-transistor *p-n-p-n* structure with regenerative action. The anode current continues to increase after some delay caused by the transit time of the injected minority carriers through the base regions. If the sum of α remains smaller than 1 when the anode current reaches I_{A1}, switching does not occur and the current approaches a stationary value asymptotically in a time interval t_m equal to the average minority-carrier lifetimes in the *n*1- and *p*2-regions. Let I_A^* be the switching current for which the sum of α equals 1, and I_{ph2} a photocurrent for which the anode current I_{A2} approaches a stationary value higher than I_A^*. Then a short time after the anode current has exceeded the value of I_A^*, the regenerative action of the feedback current will start and lead to a rapid increase of the anode current. The thyristor will switch to its *on*-state, as Fig. 26 illustrates. Note that the photocurrent I_{ph} is being amplified to I_A. The turn-on will shift to shorter time delays with increasing photocurrent, and, therefore, the turn-on time becomes shorter with enhanced light intensity.

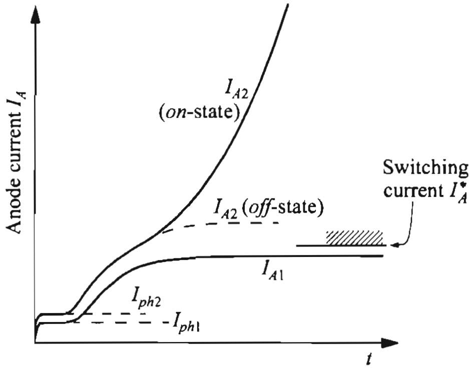


Fig. 26 Turn-on characteristics of a light-activated thyristor for two photocurrent levels. (After Ref. 34.)

A power thyristor can be turned on with very small light power (≈ 0.2 mW for a 3-kV thyristor), because the light power can be focused onto a very small area of radius r_1. For example, for a single glass fiber with 100 μm diameter, the initial turn-on area can be less than 10^{-2} mm^2. Therefore, power density in the initial turn-on area can be very high. For the cathode-shorted LASCR (Fig. 25), the minimum required light power varies approximately as r_1/r_2. Hence a smaller r_1/r_2 ratio can reduce the light power. However, even at $r_1/r_2 = 0.2$, a light power of about 5 mW is necessary for firing, which is an order of magnitude higher than an opened-cathode LASCR. Therefore, there is a trade-off between the light power and the dV/dt rating.

Once the light power turns on the initial area, the regenerative action of the device will enlarge the turned-on area, and eventually the full cathode will be on. After triggering has been brought into action and when the anode current prevails over the photocurrent, the light power can be turned off without any change of the anode current.

The amount of photocurrent needed to trigger a thyristor depends on the wavelength of the light. For silicon, the peak spectral response occurs at wavelength of 0.85 to 1.0 μm. Effective light sources with wavelength in this range include GaAs-based lasers and light-emitting diodes.

11.4 OTHER POWER DEVICES

11.4.1 Insulated-Gate Bipolar Transistor

The name insulated-gate bipolar transistor (IGBT) comes from its operation based on an internal interaction between an insulated-gate FET (IGFET) and a bipolar transistor. It has also been called by different authors as insulated-gate transistor (IGT), insulated-gate rectifier (IGR), and conductivity-modulated field-effect transistor (COMFET). The device was first demonstrated by Baliga in 1979,[35] and in 1980 by Plummer and Scharf,[36] Leipold et al.,[37] and Tihanyi.[38] A more detailed account of the device's advantages was presented in 1982 by Becke and Wheatley[39] and Baliga et al.[40] Ever since the device's conception, there have been active studies on the IGBT, with improving understanding and performance. The device has been used commercially since the late 1980s and still is gaining popularity. For more in-depth study on the device, readers are referred to Refs. 41–44.

The structure of an IGBT is shown in Fig. 27. It can be viewed as an SCR with a cathode short and an MOSFET (or more specifically, a DMOS transistor; see Section 6.5.6) connecting the n^+-cathode to the n^--base. The structure can also be viewed as a DMOS transistor with an additional *p-n* junction within the drain region. In the vertical structure (Fig. 27a), the p^+-anode is the low-resistivity substrate material, and the n^--layer is an epitaxial layer about 50 μm thick with a doping concentration below 10^{14} cm^{-3}. In this structure, isolation between devices is difficult, and devices are diced as discrete components. In the lateral structure (LIGT, lateral insulated-gate transistor) shown in Fig. 27b, the anode is incorporated at the surface, and isolation to the substrate is achieved by the *p*-type material. Like the SCR, an IGBT is made with silicon material because of its good thermal conductivity and high

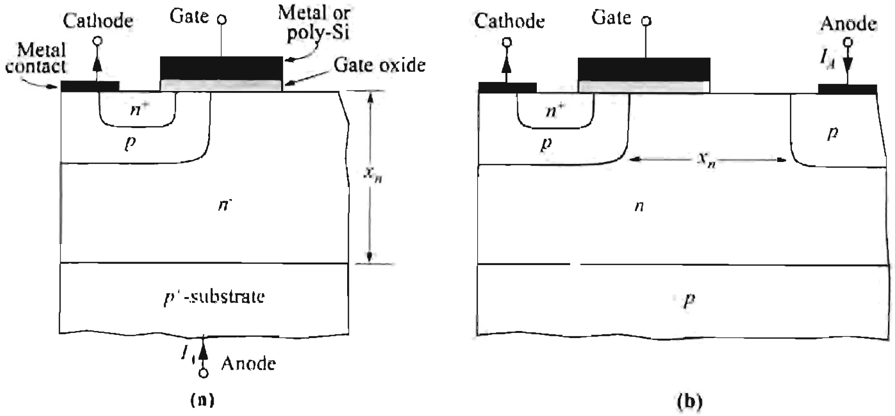


Fig. 27 (a) Vertical and (b) lateral structure of an *n*-channel IGBT.

breakdown voltage. The examples shown in Fig. 27 have an *n*-channel DMOS transistor and are called *n*-channel IGBTs. A complementary device, the *p*-channel IGBT, is also possible with opposite doping types and operated with reverse voltage polarities. The terminologies of anode/cathode/gate are adopted from an SCR, but some authors have used terms such as drain/source/gate and collector/emitter/gate.

The bulk of the device is the n^--layer, which is the drain of the DMOS transistor, as well as the base of the *p-n-p* bipolar transistor. It is lightly doped and is wide in order to support a large blocking voltage. In the *on*-state, conductivity in this region is enhanced by excess electrons injected from the n^+-cathode via the DMOS transistor surface channel, and by excess holes from the p^+-anode. This conductivity modulation is the reason for the name COMFET (conductivity-modulated FET).

With a zero gate bias, the channel of the DMOS transistor is not formed. The structure is equivalent to that of a *breakover* diode (*p-n-p-n* structure) with a cathode short (metal contact between cathode and *p*-base). The anode (or cathode) current I_A is minimal until breakdown in either polarity (Fig. 28). For forward V_{AK}, the break-

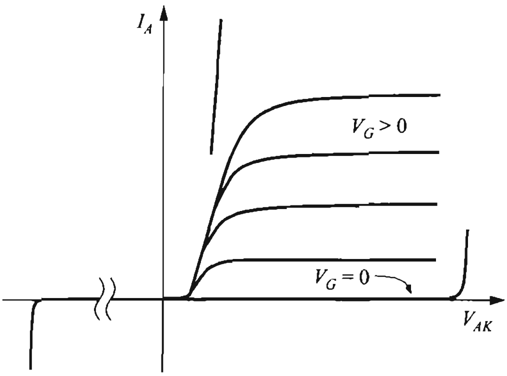


Fig. 28 Output characteristics of an *n*-channel IGBT.

down is initiated by avalanche breakdown of the n^--p junction and for reverse V_{AK}, the same process of the n^--p^+ junction. With a large positive gate voltage V_G above the threshold value, an n-type channel is induced underneath the gate, connecting the two n-type regions. Depending on the value of V_{AK}, three different modes of operation can be observed. With a small V_{AK} up to ≈ 0.7 V, the equivalent circuit is a DMOS transistor in series with a *p-i-n* diode (Fig. 29a). With negligible voltage drop across the DMOS transistor, the *p-i-n* diode is under forward bias, and the current conduction is via recombination of excess electrons and holes in the n^--region. To maintain charge neutrality, these excess electrons and holes are equal in number, being supplied from the cathode and anode, respectively. The current equation in this mode is similar to that of a *p-i-n* diode under forward bias (see Eq. 49),

$$I_A \approx \frac{4Aqn_iD_a}{x_n}\exp\left(\frac{qV_{AK}}{2kT}\right) \tag{70}$$

where A is the cross-sectional area and x_n the length of the n^--region. The factor of 2 within the exponential term is a characteristic of recombination current. This exponential rise of current with V_{AK} shows up as an offset voltage in the linear scale of Fig. 28. The current is also independent of V_G since the voltage drop across the DMOS transistor is already negligible.

The second regime starts with $V_{AK} > 0.7$ V, where the characteristics resemble those of a MOSFET. Under such medium V_{AK}, the excess holes injected from the anode cannot be totally absorbed by recombination. They spill over to the middle p-region and contribute to a *p-n-p* bipolar current. The equivalent circuit is indicated in Fig. 29b. The MOSFET current I_{MOS} becomes the base current of the bipolar transistor, and the anode current is the emitter current, given by

$$I_A = (1+\beta_{pnp})I_{MOS} \quad . \tag{71}$$

It can be seen in Fig. 28 that the anode current duplicates the general shape of the MOSFET characteristics, except amplified by a current gain. The bipolar current gain β_{pnp} is small due to the large base dimension of the n^--layer. With

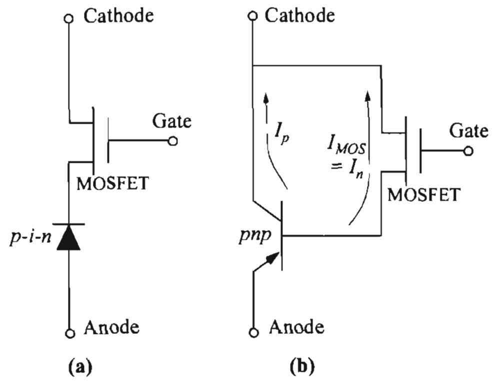


Fig. 29 Equivalent circuits of an IGBT for (a) low V_{AK} below the offset voltage, and (b) high V_{AK} above the offset voltage.

$$\beta = \frac{\alpha}{1-\alpha} \tag{72}$$

and

$$\alpha \approx \alpha_T \approx \frac{1}{\cosh(x_{nn}/L_n)} \tag{73}$$

(α_T is the base transport factor and x_{nn} is the neutral base), β_{pnp} is around unity, meaning that electron and hole currents are comparable in magnitude. Nevertheless, Eq. 71 indicates that compared to a power MOSFET of similar size, both the current and the transconductance are about doubled. This is the main feature of the IGBT.

In the third mode, if the current exceeds a critical level, the characteristics latch onto a low-impedance state, similar to the *on*-state of an SCR. This is due to the internal interaction of the *p-n-p-n* structure. In spite of the low on-resistance, this state is not desirable because once latching occurs, the gate loses control in turning off the device. The gate-controlled turn-off is critical and is precisely the advantage over an SCR. The cathode short between the n^+- and *p*-regions helps to suppress latching by decreasing the current gain of the *n-p-n* bipolar transistor. A special design with higher *p*-region concentration near the cathode has also been examined.

Besides the possibility of latching, another drawback of the IGBT is a slow turn-off process due to charge storage in the n^--region. A typical anode current waveform during turn-off is shown in Fig. 30. The decay of I_A takes place in two stages. There is first a sudden drop (ΔI_A), followed by a slow exponential decay. The initial current drop, in first-order approximation, is due to the starvation of the electron current I_n supplied by the DMOS transistor.[45] Since the current components can be separated into electron current I_n and hole current I_p, they are related to I_A by

$$I_n = (1-\alpha)I_A, \tag{74}$$

$$I_p = \alpha I_A, \tag{75}$$

and the current drop ΔI_A can be estimated ($= I_n$). The current I_p decays exponentially with a characteristic minority-carrier lifetime of holes. This turn-off process takes typically 10–50 ms and limits the IGBT to operations below 10 kHz. One technique to speed up the turn-off is to degrade the carrier lifetime by electron irradiation, but at the expense of a higher forward-voltage drop.

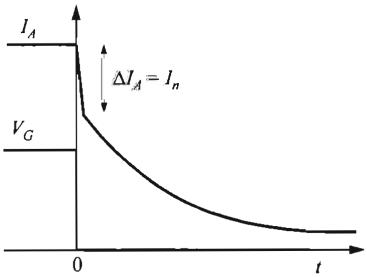


Fig. 30 Waveform of the anode–cathode current during turn-off.

The IGBT combines the salient features of a MOSFET and a bipolar transistor. In common with a MOSFET, it has high input resistance and low input capacitance. Similar to a bipolar transistor or an SCR, it has low on-resistance (or low forward voltage drop) and high-current capability. A more important feature is the gate-controlled turn-off capability. In an SCR, the gate alone cannot turn off the device, and forced commutation is needed to change the V_{AK} polarity. Such a commutation circuit adds extra cost and inflexibility. The IGBT thus has the main advantage of a gate turn-off thyristor.

11.4.2 Static-Induction Transistor

The static-induction transistor (SIT) was introduced by Nishizawa et al. in 1972.[46,47] The transistor features nonsaturating I-V characteristics with increasing drain voltage because the barrier for carriers is lowered by electrostatic induction from the drain. The static-induction transistor began to be produced in the commercial market in the mid-1980s as power amplifiers.

Structures similar, if not identical, to the SIT can be found in literature earlier, although the operations from these devices are slightly different. Shockley in 1952 proposed the analog transistor whose current is limited by space-charge-limited (SCL) current.[48] He used the name analog because of the analogy to vacuum-tube triode operation. The general characteristics of the SCL current are similar to those of the static-induction current. They both display triode-like (nonsaturating) behavior in the I_D-V_D plot as opposed to pentode-like (saturating) behavior as in a conventional FET. The SCL current is known to have a power-law dependence on drain bias, while the static-induction current has an exponential dependence. The differences will be further elaborated.

Some common structures for the static-induction transistor are shown in Fig. 31. The most critical parameters in an SIT are the spacing between gates ($2a$) and the channel doping level (N_D). Since most SITs are designed normally-on, the doping is

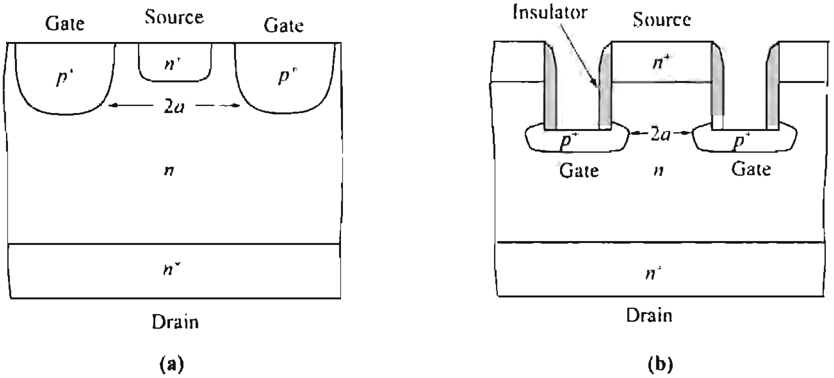


Fig. 31 Structures of static-induction transistor with (a) planar gate and (b) recessed gate.

chosen such that the depletion regions from the gates do not merge and there exists a narrow, neutral channel opening with zero gate bias. The structures also show that the gates are formed by *p-n* junctions, but the SIT operations can be generalized to include metal (Schottky) gates,[49] or MIS (metal-insulator-semiconductor) gates. In the case of metal gates, the device will be similar to a permeable-base transistor. The main difference then will be the device operation regime, not the structure. Most SITs reported are made on a Si substrate, with GaAs being the next choice of material for higher-speed operations.

The static-induction transistor is basically a JFET or MESFET with super-short channel length and with multiple gates. One major difference in structure is that the gates in the SIT do not extend close to the source or drain. As a result of the short channel (gate) length, punch-through occurs with high drain bias even if the transistor is turned off (static induction is equivalent to punch-through). The output characteristics of an SIT are shown in Fig. 32. At zero gate bias, the depletion regions around the gates do not pinch off the gap completely, and this condition corresponds to

$$\sqrt{\frac{2\varepsilon_s\psi_{bi}}{qN_D}} < a \tag{76}$$

where ψ_{bi} is the built-in potential of the *p-n* junction from the gate,

$$\psi_{bi} = \frac{kT}{q}\ln\left(\frac{N_AN_D}{n_i^2}\right) \quad . \tag{77}$$

A neutral region between the gates with zero gate bias provides a current path for a depletion-mode (normally-on) device. The current conduction is drift in nature and is similar to a buried-channel FET. With negative gate bias, the depletion regions widen and pinch off the channel, and electrons from the source start to see a potential barrier (Fig. 33). This begins when the gate voltage is more negative than

$$V_T = \psi_{bi} - V_P \tag{78}$$

while the pinch-off voltage V_P is given by

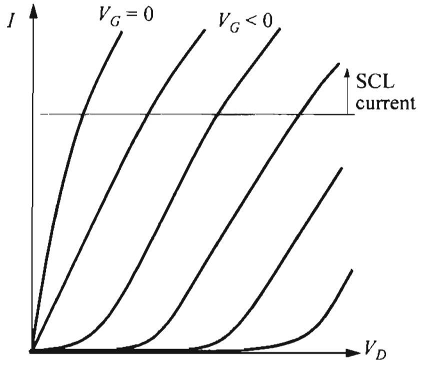


Fig. 32 Output characteristics of an SIT. At higher current levels, space-charge-limited current dominates.

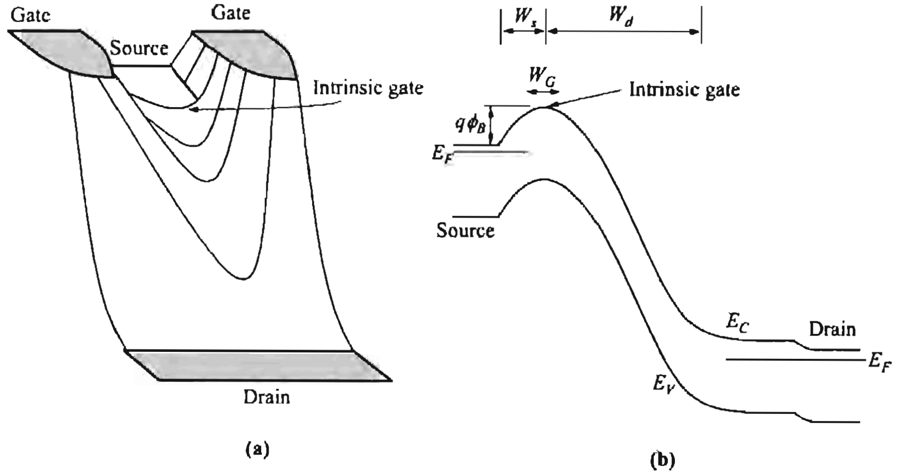

Gate
Gate
Source
Intrinsic gate
Drain
(a)
W_s
W_d
W_G
Intrinsic gate
$q\phi_B$
E_F
Source
E_C
Drain
E_F
E_V
(b)

Fig. 33 (a) Two-dimensional energy profile (conduction-band edge E_C) of an SIT. (After Ref. 50.) (b) Energy-band diagram from source to drain through the middle of the channel between gates.

$$V_P = \frac{qN_Da^2}{2\varepsilon_s} . \tag{79}$$

Once a barrier is formed, the current is controlled by diffusion, and the barrier height ϕ_B is the controlling factor for the supply of carriers from the source. This barrier height can be influenced by the gate voltage as well as the drain voltage. As shown in Fig. 34, negative gate voltage raises the barrier, and positive drain voltage lowers the barrier. The efficiency by which the terminal voltages affect the barrier is indicated by η and θ, with

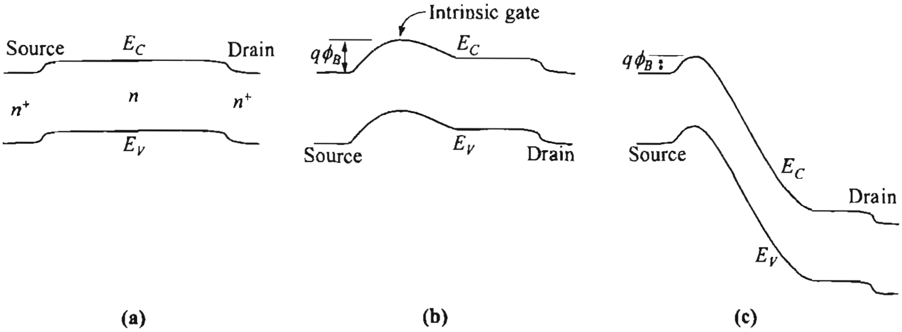

Source
E_C
Drain
n^+
n
n^+
E_V
(a)
Intrinsic gate
$q\phi_B$
E_C
Source
E_V
Drain
(b)
$q\phi_B$
Source
E_C
Drain
E_V
(c)

Fig. 34 Energy-band diagrams in the middle of the channel with biases: (a) $V_G = V_D = 0$, (b) $V_G < 0$, $V_D = 0$, barrier ϕ_B increased by negative V_G, and (c) $V_G < 0$, $V_D > 0$, barrier ϕ_B decreased by positive V_D.

$$\Delta\phi_B = -\eta\Delta V_G \tag{80}$$

and

$$\Delta\phi_B = -\theta\Delta V_D \ . \tag{81}$$

The change of barrier by the drain bias (Eq. 81) is the concept behind static induction. The factors η and θ are geometry dependent and, thus, different for different structures in Fig. 31. To use the structure in Fig. 31a as an example,[51]

$$\eta \approx \frac{W_s}{a + W_s} \ , \tag{82}$$

$$\theta \approx \frac{W_s}{W_s + W_d} , \tag{83}$$

where W_s and W_d are the depletion widths of the intrinsic gate toward the source and the drain as indicated in Fig. 33b.

The current of an SIT when the channel is pinched off is given by the form

$$J = qN_D^+\left(\frac{D_n}{W_G}\right) \exp\left(\frac{-q\phi_B}{kT}\right) \tag{84}$$

where N_D^+ is the doping concentration in the source. The term D_n/W_G is the carrier diffusion velocity. When W_G (effective thickness of the barrier, shown in Fig. 33b) becomes small, carriers are limited by the thermal velocity, giving a current of[47]

$$J = qN_D^+\sqrt{\frac{kT}{2\pi m^*}} \exp\left(\frac{-q\phi_B}{kT}\right) \ . \tag{85}$$

In either Eq. 84 or Eq. 85, the barrier height ϕ_B at the intrinsic gate is given by[52]

$$\phi_B = \frac{kT}{q}\ln\left(\frac{N_D^+}{N_D}\right) - \eta[V_G - (\psi_{bi} - V_P)] - \theta V_D \ , \quad V_G < (\psi_{bi} - V_P) . \tag{86}$$

The first term on the right is the built-in potential of the n^+-n junction, and the second and third terms are contributions from the gate and drain biases, respectively. The last term gives rise to the nonsaturating characteristics with drain bias and, thus, the static-induction effect. The channel width, pictured in Fig. 33a, is only a small fraction of the gap between gates. Since the diffusion current is exponential with ϕ_B, the effective channel width is on the order of a few Debye lengths. Such insight can be provided by computer simulations. Overall, the current can be put in the form

$$J = J_o\exp\left[\frac{q(\eta V_G + \theta V_D)}{kT}\right] \ . \tag{87}$$

At high current levels, the injected electrons are comparable to the doping level N_D. The injected carriers thus modify the field distribution, and the current is controlled by the SCL current (see Section 1.5.8). The *I-V* characteristics have the forms:

$$J = \frac{9\varepsilon_s\mu_n V_D^2}{8L^3} \ , \tag{88}$$

$$J = \frac{2\varepsilon_s \upsilon_s V_D}{L^2} \ , \tag{89}$$

$$J = \frac{4\varepsilon_s}{9L^2}\left(\frac{2q}{m^*}\right)^{1/2} V_D^{3/2} \ , \tag{90}$$

when carriers are in the mobility regime, velocity-saturation regime, or ballistic regime, respectively, and L is distance between source and drain. These equations assume that there is negligible barrier limiting the injection of carriers. In the case of an SIT, the barrier created by the gate bias controls the onset of the SCL current. In other words, SCL current starts when the ϕ_B is lowered by V_D to approximately zero. Because of this, V_D in Eqs. 88 to 90 has a threshold value and should be replaced by $(V_D + \xi V_G)$, where ξ is another constant similar in nature to η and θ.[53] With this substitution, the SCL current becomes a function of V_G. Also, comparing Eq. 88 to Eq. 87, one can now see more clearly the fundamental difference between an analog transistor and an SIT. As discussed by Nishizawa, in an analog transistor, the SCL current does not have an exponential dependence.[47] When I_D is plotted against V_D in a log-log scale, the static-induction current can have a slope higher than 2, and can be distinguished from the SCL current.

The main attractiveness of an SIT is the combination of high-voltage and high-speed capability. The low doping gives rise to high breakdown voltage up to a few hundred volts. For a buried-gate structure (not shown), the frequency of operation is limited to 2–5 MHz due to excessive parasitic capacitance. Using structures with exposed gates, the frequency can be increased to above 2 GHz. Most applications of the SIT are in the power area. As an audio power amplifier, the SIT has low noise, low distortion, and low output impedance. It can be used in high-power oscillators of microwave equipment such as communication broadcasting transmitters and microwave ovens.

Another mode of operation in the SIT family is the bipolar-mode SIT (BSIT) when the gate is forward biased to further achieve lower on-resistance.[54–55] It has also been referred to as a depleted-base transistor. In this device design the gap ($2a$) is smaller and/or the doping in the channel is lower, such that

$$\sqrt{\frac{2\varepsilon_s \psi_{bi}}{qN_D}} > a \ . \tag{91}$$

This corresponds to a pinch-off condition with zero gate bias and the device is normally-off (enhancement). With forward gate bias (positive), the barrier is lowered since the built-in potential is reduced. Furthermore, the p^+-gates, being forward biased, inject holes to the channel. The holes get collected at a potential minimum (energy maximum) at the intrinsic gate, raise the potential, and enhance the electron supply from the source. This mode of operation is similar to a bipolar transistor except that here the intrinsic gate is a virtual base whose potential is accessed by the p^+-gate (or base in bipolar terminology) indirectly. At this point, the electron concentration is much higher than the background doping level, so the current is larger than a conventional JFET. The output characteristics of a BSIT are shown in Fig. 35a.

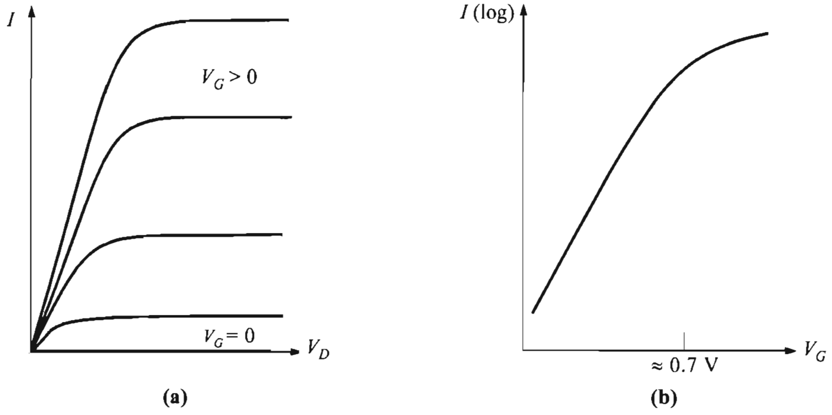


Fig. 35 (a) Output characteristics of a BSIT. (b) For a fixed V_D, the current rises exponentially with V_G, similar to the base-emitter bias of a bipolar transistor or subthreshold current of an FET, until the gate diode is too strongly biased (> 0.7 V).

They are drastically different from an SIT in that the currents are saturating with V_D (pentode-like rather than triode-like). The general characteristics resemble those of a bipolar transistor.

11.4.3 Static-Induction Thyristor

The static-inductor thyristor (SIThy) is also called a field-controlled thyristor. Over a large portion of the operating regime, the device is similar to the static-induction transistor which was conceived around the same time. The static-induction thyristor was presented in part of a paper by Nishizawa et al.[47] and was described in more details by Houston et al.,[56] both in 1975.

The basic structure of the static-induction thyristor, shown in Fig. 36, is a *p-i-n* diode with part of the channel surrounded by closely spaced junction gates (or grids). It is also similar to the SIT with the p^+-anode replacing the n^+-drain. The structure can be made in the form of a planar gate or a buried gate. The advantage of the planar gate is a lower gate resistance since a metal contact can be deposited directly over it. This results in a smaller gate debiasing effect during turn-off when there is substantial current going through the gate. The advantages of the buried gate are a more-efficient use of cathode area, and more effective gate control of the current, resulting in a higher forward-blocking voltage gain (to be discussed later). The double-gate SIThy has been shown to be capable of higher speed and lower voltage drop than in single-gate structures.[57]

In a static-induction thyristor, the gate controls the current in two distinct ways. Using the structure in Fig. 36b as an example, before pinch-off (Fig. 37a), the depletion regions of the two gates do not merge, and the gate voltage controls the effective cross-sectional area of the *p-i-n* diode between the anode and the cathode. For a large negative gate voltage, the junctions are under reverse bias and the depletion regions

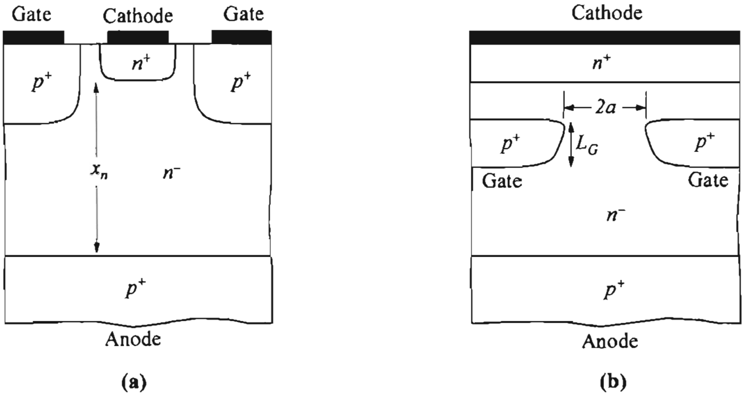

Gate
Cathode
Gate
n^+
p^+
p^+
x_n
n^-
p^+
Anode
(a)
Cathode
n^+
$2a$
p^+
L_G
p^+
Gate
Gate
n^-
p^+
Anode
(b)

Fig. 36 Structures of the static-inductor thyristor with (a) planar gates and (b) buried gates (grids).

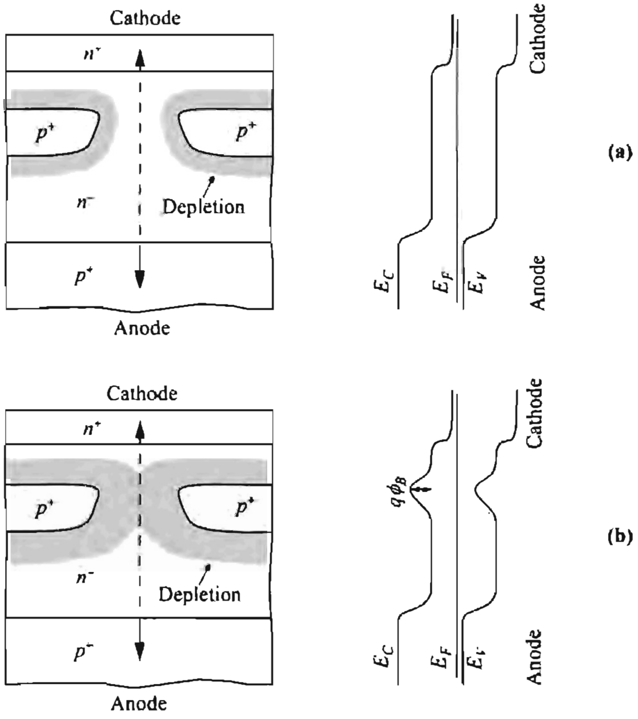

Cathode
n^+
p^+
p^+
n^-
Depletion
p^+
Anode
E_C
E_F
E_V
Anode
Cathode
(a)
Cathode
n^+
p^+
p^+
n^-
Depletion
p^+
Anode
$q\phi_B$
E_C
E_F
E_V
Anode
Cathode
(b)

Fig. 37 Schematic diagrams showing the effects of depletion width on the channel, and their energy-band diagrams at zero V_{AK}. (a) Before pinch-off. (b) After pinch-off. Energy-band diagram is along the middle of the channel shown by the dashed line.

widen and eventually meet (Fig. 37b). Under this pinch-off condition, a barrier for electrons is formed that controls the current flow.

The gate voltage corresponding to pinch-off can be approximated by a simple one-dimensional depletion theory,

$$V_P = \psi_{bi} - \frac{qN_D a^2}{2\varepsilon_s} \tag{92}$$

where ψ_{bi} is the built-in potential of the gate junction. By adjusting the gap $2a$ between gates, one can design a device to be normally-on or normally-off. In a normally-on SIThy, pinch-off does not occur with zero gate voltage, and a high current can flow. In a normally-off SIThy, $2a$ is smaller (or N_D in the n^--layer is lower) such that pinch-off occurs with zero gate bias. In order to turn on the device, the gate has to be forward biased to reduce the depletion regions to open up a channel. The normally-off device is less common, due to the larger gate current under forward bias.

The output characteristics for a normally-on SIThy are shown in Fig. 38. Before pinch-off, the current conduction is that of a *p-i-n* diode, given by (see Eq. 49)

$$I_A = \frac{4AqD_a n_i}{x_n} \exp\left(\frac{qV_{AK}}{2kT}\right) \tag{93}$$

which is a recombination current of excess electrons and holes in the n^--region. D_a is the ambipolar diffusion coefficient. Under forward bias V_{AK}, electrons are injected from the cathode and holes from the anode, and their concentrations are equal to maintain charge neutrality. These excess electrons and holes increase the conductivity of the n^--layer. This phenomenon is called conductivity modulation. Note that although the output characteristics are similar in shape to those of the SIT, the p^+-anode can inject holes and enable conductivity modulation, resulting in a lower forward-voltage drop or lower on-resistance.

With a larger reverse gate bias, pinch-off is reached and a barrier for electrons is formed (Fig. 37b). This barrier limits the electron supply and becomes the controlling factor for the overall current. Without an ample electron supply, the hole current

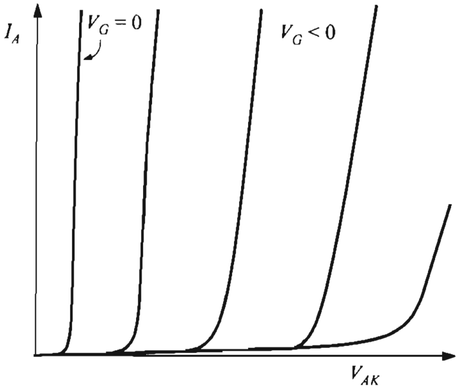


Fig. 38 Output characteristics of a normally-on static-induction thyristor. For a normally-off device, similar curves are obtained with forward gate voltages.

reduces to the diffusion current and becomes insignificant. The barrier height ϕ_B not only is controlled by the gate voltage, it can also be lowered by a large V_{AK}. This dependence of ϕ_B on V_{AK}, called static induction, is the main current conduction mechanism in a static-induction transistor. Static-induction current is basically a punch-through current due to the thin barrier in the direction of the current flow. It is a diffusion current with the barrier controlling the supply of carriers, given in the form

$$\begin{aligned} I_A &\propto \exp\left(\frac{-q\phi_B}{kT}\right) \\ &= I_{o2}\exp\left[\frac{q(\eta V_G+\theta V_{AK})}{kT}\right] . \end{aligned} \tag{94}$$

η and θ indicate the control of V_G and V_{AK} on the barrier height.

One useful parameter for the SIThy is the forward-blocking voltage gain μ, which is defined as the change of V_{AK} induced by the change of V_G for the same anode current. According to the previous expression, it is equal to

$$\mu = -\left.\frac{dV_{AK}}{dV_G}\right|_{I_A} = \frac{\eta}{\theta} . \tag{95}$$

It has been shown experimentally that[58]

$$\mu \approx \frac{4L_G W_d}{a^2} \tag{96}$$

where W_d is the depletion width of the gate junction in the direction toward the anode (Fig. 33b).

One of the advantages of the SIThy is higher-speed operation than in an SCR, due to a faster turn-off process. During turn-off, the reverse gate bias can extract the excess minority carriers (holes) quickly. The excess electrons, being majority carriers in the n^--layer, can be swept off quickly by the drift process. The hole current contributes to an instantaneously large gate current, and a small gate resistance is critical to avoid gate debiasing. An alternative technique to reduce the turn-off time is to reduce the minority-carrier lifetime by proton or electron irradiation. The penalty for using this technique is a larger forward voltage drop.

It has also been proposed to use light to trigger or quench a SIThy.[59] When a SIThy is off, either as a normally-off device or turned off by a gate bias, light-generated holes get trapped at the barrier (Fig. 37b). These positive charges decrease the barrier height for electrons and trigger turn-on. For a normally-on SIThy, the gate is connected to a negative voltage source via a phototransistor. Light can activate the phototransistor, and the negative voltage source is applied to the gate to turn off the SIThy.

The static-induction thyristor offers certain advantages over other thyristors. Due to the faster turn-off, higher operating frequency is possible. Because the turn-on process does not depend on regenerative feedback as in an SCR, it has more stable operation at higher temperatures and can tolerate faster dI/dt and dV/dt transients. It has low forward-voltage drop, high blocking-voltage gain up to ≈ 700, and gate-controlled turn-off capability (an SCR after latching cannot be turned off simply by removing the gate bias). The SIThy has been applied mainly in power source conversion such as ac-to-dc converters, dc-to-ac converters, and chopper circuits.[60] Another application is pulse generation, for induction heating, lighting of fluorescent lamps, and driving pulsed lasers.

REFERENCES

1. W. Shockley, *Electrons and Holes in Semiconductors*, D. Van Nostrand, Princeton, New Jersey, 1950, p. 112.
2. J. J. Ebers, "Four-Terminal *p-n-p-n* Transistors," *Proc. IRE*, **40**, 1361 (1952).
3. J. L. Moll, M. Tanenbaum, J. M. Goldey, and N. Holonyak, "*p-n-p-n* Transistor Switches," *Proc. IRE*, **44**, 1174 (1956).
4. I. M. Mackintosh, "The Electrical Characteristics of Silicon *p-n-p-n* Triodes," *Proc. IRE*, **46**, 1229 (1958).
5. R. W. Aldrich and N. Holonyak, Jr., "Multiterminal *p-n-p-n* Switches," *Proc. IRE*, **46**, 1236 (1958).
6. F. E. Gentry, F. W. Gutzwieler, N. H. Holonyak, and E. E. Von Zastrow, *Semiconductor Controlled Rectifiers*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.
7. A. Blicher, *Thyristor Physics*, Springer, New York, 1976.
8. S. K. Ghandhi, *Semiconductor Power Devices*, Wiley, New York, 1977.
9. B. J. Baliga, *Power Semiconductor Devices*, PWS, Boston, 1996.
10. P. D. Taylor, *Thyristor Design and Realization*, Wiley, New York, 1987.
11. S. M. Sze and G. Gibbons, "Avalanche Breakdown Voltages of Abrupt and Linearly Graded *p-n* Junctions in Ge, Si, GaAs, and GaP," *Appl. Phys. Lett.*, **8**, 111 (1966).
12. A. Herlet, "The Maximum Blocking Capability of Silicon Thyristors," *Solid-State Electron.*, **8**, 655 (1965).
13. E. E. Haller, "Isotopically Engineered Semiconductors," *J. Appl. Phys.*, **77**, 2857 (1995).
14. E. W. Haas and M. S. Schnoller, "Phosphorus Doping of Silicon by Means of Neutron Irradiation," *IEEE Trans. Electron Dev.*, **ED-23**, 803 (1976).
15. J. Cornu, S. Schweitzer, and O. Kuhn, "Double Positive Beveling: A Better Edge Contour for High Voltage Devices," *IEEE Trans. Electron Dev.*, **ED-21**, 181 (1974).
16. R. L. Davies and F. E. Gentry, "Control of Electric Field at the Surface of *p-n* Junctions," *IEEE Trans. Electron Dev.*, **ED-11**, 313 (1964).
17. F. E. Gentry, "Turn-on Criterion for *p-n-p-n* Devices," *IEEE Trans. Electron Dev.*, **ED-11**, 74 (1964).
18. E. S. Yang and N. C. Voulgaris, "On the Variation of Small-Signal Alphas of a *p-n-p-n* Device with Current," *Solid-State Electron.*, **10**, 641 (1967).
19. A. Munoz-Yague and P. Leturcq, "Optimum Design of Thyristor Gate-Emitter Geometry," *IEEE Trans. Electron Dev.*, **ED-23**, 917 (1976).
20. M. S. Adler, "Accurate Calculation of the Forward Drop and Power Dissipation in Thyristors," *IEEE Trans. Electron Dev.*, **ED-25**, 16 (1978).

21. H. F. Storm and J. G. St. Clair, "An Involute Gate-Emitter Configuration for Thyristors," *IEEE Trans. Electron Dev.*, **ED-21**, 520 (1974).
22. F. E. Gentry and J. Moyson, "The Amplifying Gate Thyristor," Paper No. 19.1, *IEEE Meet. Prof. Group Electron Devices*, Washington, D.C., 1968.
23. J. F. Gibbons, "Graphical Analysis of the *I-V* Characteristics of Generalized *p-n-p-n* Devices," *Proc. IEEE*, **55**, 1366 (1967).
24. E. S. Yang, "Turn-off Characteristics of *p-n-p-n* Devices," *Solid-State Electron.*, **10**, 927 (1967).
25. T. S. Sundresh, "Reverse Transient in *p-n-p-n* Triodes," *IEEE Trans. Electron Dev.*, **ED-14**, 400 (1967).
26. B. J. Baliga and E. Sun, "Comparison of Gold, Platinum, and Electron Irradiation for Controlling Lifetime in Power Rectifiers," *IEEE Trans. Electron Dev.*, **ED-24**, 685 (1977).
27. B. J. Baliga and S. Krishna, "Optimization of Recombination Levels and their Capture Cross Section in Power Rectifiers and Thyristors," *Solid-State Electron.*, **20**, 225 (1977).
28. J. Shimizu, H. Oka, S. Funakawa, H. Gamo, T. Ilda, and A. Kawakami, "High-Voltage High-Power Gate-Assisted Turn-Off Thyristor for High-Frequency Use," *IEEE Trans. Electron Dev.*, **ED-23**, 883 (1976).
29. E. Schlegel, "Gate Assisted Turn-off Thyristors," *IEEE Trans. Electron Dev.*, **ED-23**, 888 (1976).
30. F. M. Roberts and E. L. G. Wilkinson, "The Relative Merits of Thyristors and Power Transistors for Fast Power-Switching Application," *Int. J. Electron.*, **33**, 319 (1972).
31. E. D. Wolley, "Gate Turn-Off in *p-n-p-n* Devices," *IEEE Trans. Electron Dev.*, **ED-13**, 590 (1966).
32. F. E. Gentry, R. I. Scace, and J. K. Flowers, "Bidirectional Triode *p-n-p-n* Switches," *Proc. IEEE*, **53**, 355 (1965).
33. J. F. Essom, "Bidirectional Triode Thyristor Applied Voltage Rate Effect Following Conduction," *Proc. IEEE*, **55**, 1312 (1967).
34. W. Gerlach, "Light Activated Power Thyristors," *Inst. Phys. Conf. Ser.*, **32**, 111 (1977).
35. B. J. Baliga, "Enhancement- and Depletion-Mode Vertical-Channel M.O.S. Gated Thyristors," *Electron. Lett.*, **15**, 645 (1979).
36. J. D. Plummer and B. W. Scharf, "Insulated-Gate Planar Thyristors: I–Structure and Basic Operation," *IEEE Trans. Electron Dev.*, **ED-27**, 380 (1980).
37. L. Leipold, W. Baumgartner, W. Ladenhauf, and J. P. Stengl, "A FET-Controlled Thyristor in SIPMOS Technology," *Tech. Dig. IEEE IEDM*, 79 (1980).
38. J. Tihanyi, "Functional Integration of Power MOS and Bipolar Devices," *Tech. Dig. IEEE IEDM*, 75 (1980).
39. H. W. Becke and C. F. Wheatley, Jr., "Power MOSFET with an Anode Region," U.S. Patent 4,364,073 (1982).
40. B. J. Baliga, M. S. Adler, P. V. Gray, R. P. Love, and N. Zommer, "The Insulated Gate Rectifier (IGR): A New Power Switching Device," *Tech. Dig. IEEE IEDM*, 264 (1982).
41. V. K. Khanna, *The Insulated Gate Bipolar Transistor (IGBT): Theory and Design*, Wiley/IEEE Press, Hoboken, New Jersey, 2003.
42. A. R. Hefner, Jr. and D. L. Blackburn, "An Analytical Model for the Steady-State and Transient Characteristics of the Power Insulated-Gate Bipolar Transistor," *Solid-State Electron.*, **31**, 1513 (1988).

43. D. S. Kuo, C. Hu, and S. P. Sapp, "An Analytical Model for the Power Bipolar-MOS Transistor," *Solid-State Electron.*, **29**, 1229 (1986).
44. H. Yilmaz, W. Ron Van Dell, K. Owyang, and M. F. Chang, "Insulated Gate Transistor Physics: Modeling and Optimization of the On-State Characteristics," *IEEE Trans. Electron Dev.*, **ED-32**, 2812 (1985).
45. B. J. Baliga, "Analysis of Insulated Gate Transistor Turn-Off Characteristics," *IEEE Electron Dev. Lett.*, **EDL-6**, 74 (1985).
46. J. Nishizawa, "A Low Impedance Field Effect Transistor," *Tech. Dig. IEEE IEDM*, 144 (1972).
47. J. I. Nishizawa, T. Terasaki, and J. Shibata, "Field-Effect Transistor Versus Analog Transistor (Static Induction Transistor)," *IEEE Trans. Electron Dev.*, **ED-22**, 185 (1975).
48. W. Shockley, "Transistor Electronics: Imperfections, Unipolar and Analog Transistors," *Proc. IRE*, **40**, 1289 (1952).
49. P. M. Campbell, W. Garwacki, A. R. Sears, P. Menditto, and B. J. Baliga, "Trapezoidal-Groove Schottky-Gate Vertical Channel GaAs FET (GaAs Static Induction Transistor)," *Tech. Dig. IEEE IEDM*, 186 (1984).
50. J. I. Nishizawa and K. Yamamoto, "High-Frequency High-Power Static Induction Transistor," *IEEE Trans. Electron Dev.*, **ED-25**, 314 (1978).
51. J. I. Nishizawa, "Junction Field-Effect Devices," *Proc. Brown Boveri Symp.*, 241 (1982).
52. C. Bulucea and A. Rusu, "A First-Order Theory of the Static Induction Transistor," *Solid-State Electron.*, **30**, 1227 (1987).
53. O. Ozawa and K. Aoki, "A Multi-Channel FET with a New Diffusion Type Structure," *Jpn. J. Appl. Phys.*, Suppl., **15**, 171 (1976).
54. J. I. Nishizawa, T. Ohmi, and H. L. Chen, "Analysis of Static Characteristics of a Bipolar-Mode SIT (BSIT)," *IEEE Trans. Electron Dev.*, **ED-29**, 1233 (1982).
55. T. Tamama, M. Sakaue, and Y. Mizushima, "'Bipolar-Mode' Transistors on a Voltage-Controlled Scheme," *IEEE Trans. Electron Dev.*, **ED-28**, 777 (1981).
56. D. E. Houston, S. Krishna, D. Piccone, R. J. Finke, and Y. S. Sun, "Field Controlled Thyristor (FCT)–A New Electronic Component," *Tech. Dig. IEEE IEDM*, 379 (1975).
57. J. Nishizawa, Y. Yukimoto, H. Kondou, M. Harada, and H. Pan, "A Double-Gate-Type Static-Induction Thyristor," *IEEE Trans. Electron Dev.*, **ED-34**, 1396 (1987).
58. J. Nishizawa, K. Muraoka, T. Tamamushi, and Y. Kawamura, "Low-Loss High-Speed Switching Devices, 2300-V 150-A Static Induction Thyristor," *IEEE Trans. Electron Dev.*, **ED-32**, 822 (1985).
59. J. Nishizawa, T. Tamamushi, and K. Nonaka, "Totally Light Controlled Static Induction Thyristor," *Physica*, **129B**, 346 (1985).
60. J. Nishizawa, "Application of the Power Static Induction (SI) Devices," *Proc. PCIM*, 1 (1988).

PROBLEMS

1. For the doping profile shown in Fig. 1b, find the width of the $n1$-region so that the thyristor has a reverse blocking voltage of 200 V.

2. If we use SiC for power devices, estimate the maximum blocking voltage and the required minimum $n1$-layer thickness for the doping profile shown in Fig. 1b, assuming the $p1$-$n1$-$p2$ bipolar common-base current gain is very small.
3. Compare the doping variation (as a percentage of the average value) of the conventionally doped silicon with the neutron-transmutation doped sample shown in Fig. 4.
4. If the current gain α_2 for the $n1$-$p2$-$n2$ transistor is 0.4 independent of current, and α_1 of the $p1$-$n1$-$p2$ transistor can be expressed as $0.5\sqrt{L_p/W}\ln(J/J_0)$, where L_p is 25 µm, $W = 40$ µm, and J_0 is 5×10^{-6} A/cm^2, find the cross-sectional area of the thyristor that will switch at a current I_s of 1 mA.
5. With a cathode short, the current gain α_2 of the $n1$-$p2$-$n2$ transistor in the thyristor is degraded to α_2' as given by Eq. 20. Derive the equation.
6. In a silicon thyristor, the $p1$-$n1$-$p2$ section is similar to a p^+-i-n^+ diode since the $n1$ doping is very low. (a) Estimate the voltage drop across the i-region, (b) the carrier density in the i-region, and (c) the effective resistance of the i-region for a forward conduction current of 200 A/cm^2. The mobility ratio $b \equiv \mu_n/\mu_p$ is assumed to be 3 independent of carrier concentration, the $n1$-layer is 50 µm thick, the effective lifetime is 10^{-6} s, and the device cross-sectional area is 1 cm^2.
7. The silicon thyristor, shown is Fig. 1b, has W_{n1} ($n1$-layer thickness) of 50 µm and W_{p2} ($p2$-layer thickness) of 10 µm. The doping in $n1$-region is 10^{14} cm^{-3} and the doping in $p2$-region is assumed to be a constant = 10^{17} cm^{-3}. If the holding current is 0.1 A, the forward-conduction current I_F is 10 A, and the lifetime in the $n1$-layer is 10^{-7} s, find the turn-on and turn-off times.
8. A silicon gate turn-off thyristor has doping concentrations of 10^{14} cm^{-3} and 10^{17} cm^{-3} for the $n1$- and $p2$-regions, respectively. The thickness of $n1$-region is 100 µm, and $n2$ is 10 µm. If the device is operated at an anode current of 100 A, find the minimum negative gate current required to turn off the device. The minority-carrier lifetime is 0.15 µs in the $n1$-region and 4 µs in the $p2$-region.
9. Determine the doping concentration and thickness of the drift region for a symmetric blocking silicon n-channel IGBT structure with a breakdown voltage of 500 V if the lifetime in the drift region is 1 µs.
   (*Hint*: As a general guideline, the thickness of the drift region is chosen so that it is equal to the depletion width at the maximum operating voltage plus one diffusion length.)
10. A silicon IGBT, as shown in Fig. 27a, has a channel length of 3 µm, a channel width of 16 µm, a p-base doping of 1×10^{17} cm^{-3}, a gate oxide of 0.02 µm, an $n1$-region thickness of 70 µm, an anode area of 16 µm×16 µm, a lifetime of 1 µs in the $n1$-region, and a channel mobility of 500 cm^2/V-s. Calculate the on-state voltage drop for this IGBT with a current density of 200 A/cm^2 and $(V_G - V_T) = 5$ V.

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

PART V

PHOTONIC DEVICES AND SENSORS

- **Chapter 12 LEDs and Lasers**
- **Chapter 13 Photodetectors and Solar Cells**
- **Chapter 14 Sensors**

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

12

LEDs and Lasers

12.1 INTRODUCTION
12.2 RADIATIVE TRANSITIONS
12.3 LIGHT-EMITTING DIODE (LED)
12.4 LASER PHYSICS
12.5 LASER OPERATING CHARACTERISTICS
12.6 SPECIALTY LASERS

12.1 INTRODUCTION

Photonic devices are those in which the basic particle of light—the photon, plays a major role. Photonic devices can be divided into three groups: (1) devices as light sources that convert electrical energy into optical radiation—the LED (*light-emitting diode*) and the diode laser (*light amplification by stimulated emission of radiation*), (2) devices that detect optical signals—photodetectors, and (3) devices that convert optical radiation into electrical energy—the photovoltaic device or solar cell. The first group is considered in this chapter; photodetectors and solar cell are discussed in Chapter 13.

The electroluminescence phenomenon was discovered in 1907.[1] Electroluminescence is the generation of light by an electric current passing through a device under bias. Electroluminescent light differs from thermal radiation (incandescence) in the relatively narrow range of wavelengths contained within its spectrum. For LEDs, the spectral line width is typically 5 to 20 nm. The light may even be nearly perfectly monochromatic as in the laser diode, with a line width of 0.1 to 1 Å. LEDs and lasers are the only light sources from semiconductor devices. As shown in this chapter, they play an increasingly important role in our daily lives, as well as pushing many frontiers of science such as in communication and medicine.

The LED and semiconductor laser belong to the luminescent device family. Luminescence is the emission of optical radiation (ultraviolet, visible, or infrared) as a result of electronic excitation in a device or material, excluding any radiation that is purely the result of the temperature of the material (incandescence). Figure 1 shows a chart of the visible and near-visible portions of the electromagnetic spectrum. Although different methods may be used to excite radiations of different wavelengths, all radiations are fundamentally alike. The visual range of the human eye extends only from about 0.4 to 0.7 μm. Figure 1 shows the major color bands from violet to red. The infrared region extends from 0.7 to about 1000 μm, and the ultraviolet region includes wavelengths from 0.4 to about 0.01 μm (i.e., 10 nm). In this and subsequent chapters, we are primarily interested in the wavelengths ranging from near ultraviolet (≈ 0.3 μm) to near infrared (≈ 1.5 μm).

The effectiveness of light for stimulating the human eye is given by the relative eye sensitivity [or luminous efficiency $V(\lambda)$] which is a strong function of the wavelength. Figure 1 shows the relative eye sensitivity, for a 2° viewing angle, as defined by the Commission Internationale de l'Eclairage (CIE) for photopic vision.[2] For the maximum sensitivity of the eye at 0.555 μm, $V(0.555\ \mu\text{m}) = 1.0$; the value of $V(\lambda)$ falls to nearly zero at the boundaries of the visible spectrum at 0.4 and 0.7 μm. So for the colors of red and violet, the eye sensitivity is lower compared to green, and it takes higher intensity to achieve similar brightness as observed by the human eyes.

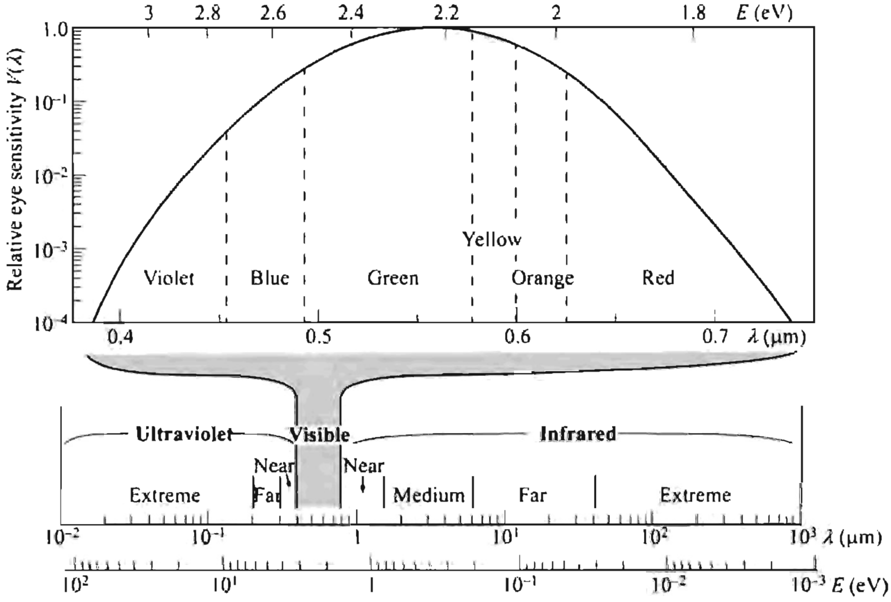


Fig. 1 Visible and near-visible electromagnetic spectrum. The visible portion is expanded at the top, and divided into major color bands. Also indicated is relative luminosity function $V(\lambda)$ as defined by the CIE for normal photopic vision.

12.2 RADIATIVE TRANSITIONS

Figure 2 schematically demonstrates the basic recombination transitions of excess carriers in a semiconductor. These transitions may be classified as follows. The first classification [label (1)] is the interband transition: (a) intrinsic emission corresponding very closely in energy to the bandgap, and (b) higher-energy emission involving energetic or hot carriers, sometimes related to avalanche emission. The second classification (2) is the transitions involving chemical impurities or physical defects; (a) conduction band to acceptor-type defect, (b) donor-type defect to valence band, (c) donor-type to acceptor-type defects (pair emission), and (d) band-to-band via deep-level traps. The third classification (3) is the intraband transition involving hot carriers, sometimes called deceleration emission or Auger process. Not all transitions can occur in the same material or under the same conditions, and not all transitions are radiative. An efficient luminescent material is one in which radiative transitions predominate over nonradiative ones such as the Auger nonradiative recombination where the band-to-band recombination energy is transferred to a hot electron or hole excited within a band.[2] In comparison, it will be shown that band-to-band recombination [(a) in (1)] is the most probable radiative process.

12.2.1 Emission Spectra

There are three main optical processes for interaction between a photon and an electron in a solid (Fig. 3): (a) A photon may be absorbed by the excitation of an electron from a filled state in the valence band to an empty state in the conduction band. (b) An electron in the conduction band can spontaneously return to an empty state in the valence band (recombination), with the emission of a photon. This process (b) is thus

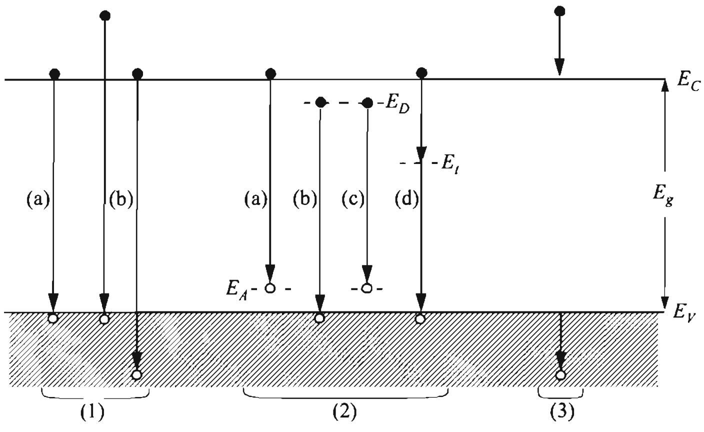


Fig. 2 Basic recombination transitions in semiconductor. E_D, E_A, E_t are donor-type, acceptor-type, and deep-level traps respectively. (After Ref. 3.)

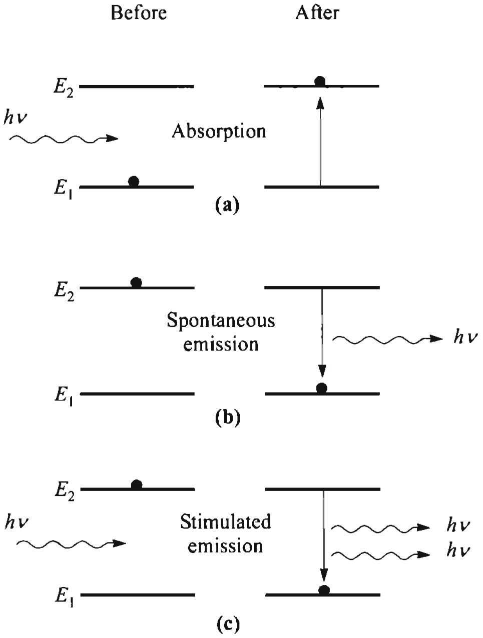


Fig. 3 The three basic optical processes between two energy levels. The black dot indicates the state of the electron. The initial state is at the left; the final state, after the process has occurred, is at the right.

the reverse of process (a), and (c) the incoming photon can stimulate the emission of another similar photon by recombination, giving out a net of two photons which are coherent. Process (a) is the main process in a photodetector or solar cell, process (b) is the main process in an LED, while process (c) is the process in a laser.

Either for photon absorption or emission, the conventional theory for optical transitions between the valence and conduction bands of direct-bandgap materials is based on the so-called $\boldsymbol{k}$-selection rule. From conservation of momentum, the wave vector $\boldsymbol{k}_1$ of the valence-band wave function and the wave vector $\boldsymbol{k}_2$ of the conduction-band must differ by the wave vector of the photon. Since the wave vector of the electron is much larger than that of the photon, the $\boldsymbol{k}$-selection rule is generally written as

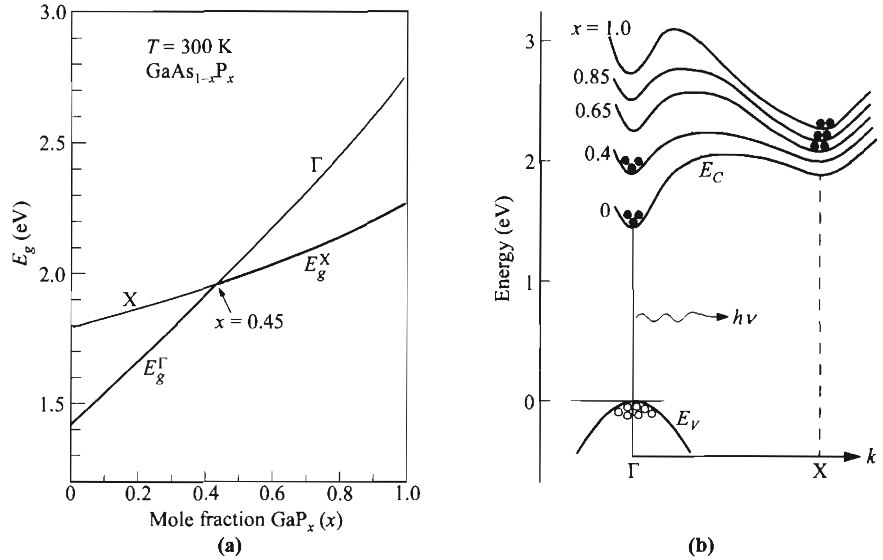
$$\boldsymbol{k}_1 = \boldsymbol{k}_2 . \tag{1}$$

The allowed transitions are then between initial and final states of the same wave vector and are called *direct* or *vertical* transitions (in E-k space).

When the conduction-band minima are not at the same value of $\boldsymbol{k}$ as the valence band, assistance of a phonon is necessary to conserve crystal momentum, and the transition is called *indirect*. Radiative transitions in indirect-bandgap materials are much less probable. To encourage light emission in indirect-bandgap semiconductors, special types of impurities are introduced. The wave functions change and the *k*-selection rule does not hold.

Figure 4a shows the energy gap for $GaAs_{1-x}P_x$ as a function of the mole fraction x. For $0 < x < 0.45$, the energy gap is direct and increasing from $E_g = 1.424$ eV at $x = 0$ to $E_g = 1.977$ eV at $x = 0.45$. For $x > 0.45$, the energy gap is indirect. Figure 4b shows the corresponding energy-momentum plots for some selected alloy compositions. As indicated, the conduction band has two minima. The one along the Γ-axis is the direct minimum, whereas the one along the X-axis is the indirect minimum. Electrons in the direct minimum of the conduction band and holes at the top of the valence band have equal momentum; while electrons in the indirect minimum have different momentum. For direct-bandgap semiconductors, such as GaAs and $GaAs_{1-x}P_x$ ($x \leq 0.45$), the momentum is conserved and interband transitions may occur with high probability. The photon energy is then approximately equal to the bandgap energy of the semiconductor. The radiative transition mechanism is predominant in direct-bandgap materials. However, for $GaAs_{1-x}P_x$ with $x > 0.45$ and GaP that are indirect-bandgap semiconductors, the probability for interband transitions is extremely small, since phonons or other scattering agents must participate in the process in order to conserve momentum. Therefore, for indirect-bandgap semiconductors, special type of recombination centers are incorporated to enhance the radiative transition.

So far we have assumed that the band-to-band recombination corresponds to the energy of the bandgap. In practice, electrons and holes reside slightly above and below their band edges E_C and E_V respectively, due to thermal energy when the tem-



(a)

(b)

Fig. 4 (a) Composition dependence of the direct and indirect bandgap for $GaAs_{1-x}P_x$. (After Ref. 4.) (b) Energy-momentum diagram for $GaAs_{1-x}P_x$. (After Ref. 5.)

perature is above absolute zero. As a result, the photon energy emitted would be slightly larger than the bandgap energy. We analyze here the spectrum of the spontaneous emission. Near the band edges, the energy of the emitted photon is governed by the relationship

$$\begin{aligned} h\nu &= \left(E_C + \frac{\hbar^2 k^2}{2m_e^*}\right) - \left(E_V - \frac{\hbar^2 k^2}{2m_h^*}\right) \\ &= E_g + \frac{\hbar^2 k^2}{2m_r^*}\ . \end{aligned} \tag{2}$$

The above is called the joint dispersion relation and m_r^* is the reduced effective mass

$$\frac{1}{m_r^*} = \frac{1}{m_e^*} + \frac{1}{m_h^*}\ . \tag{3}$$

With similar treatment, a joint density of states can be obtained as[6]

$$N_J(E) = \frac{(2m_r^*)^{3/2}}{2\pi^2\hbar^3}\sqrt{E - E_g}\ . \tag{4}$$

The distribution of carriers is governed by the Boltzmann distribution

$$F(E) = \exp\left(-\frac{E}{kT}\right)\ . \tag{5}$$

The spontaneous emission rate is proportional to the product of Eqs. 4 and 5, and it generally has the form[6]

$$I(E{=}h\nu) \propto \sqrt{E - E_g}\exp\left(-\frac{E}{kT}\right)\ . \tag{6}$$

The essence of Eq. 6 is depicted in Fig. 5. The spectrum of spontaneous emission has a threshold energy of E_g, a peak of ($E_g + \frac{1}{2}kT$), and a half-power width of $1.8kT$. This translates into a spectrum width in wavelength of

$$\Delta\lambda \approx \frac{1.8kT\lambda^2}{hc}\ , \tag{7}$$

where c is the velocity of light. In the middle of the visible spectrum, the emission spectrum width is around 10 nm.

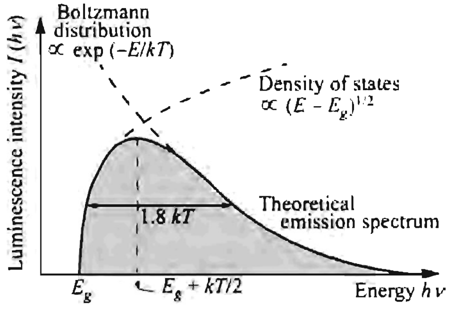


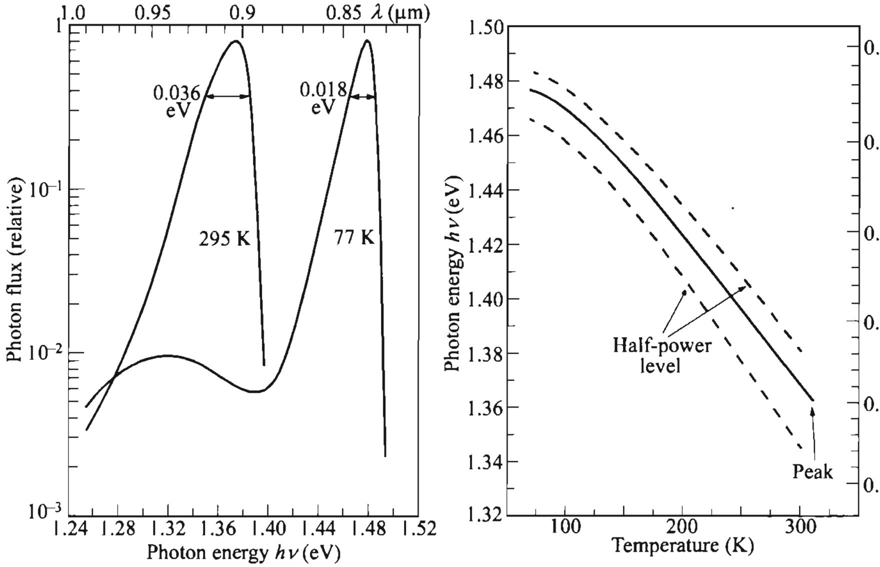
Fig. 5 Theoretical spectrum of spontaneous emission. (After Ref. 6).

Figure 6a shows the emission spectra for a GaAs *p-n* junction observed at 77 and 300 K. The peak photon energy decreases with increasing temperature mainly because the bandgap decreases. Figure 6b shows a more-detailed plot for the peak photon energy and the half-power points from the diode emission spectrum as a function of temperature. The width of the half-power points increases slightly with temperature, as expected from Eq. 6.

12.2.2 Methods of Excitation

The types of luminescence may be distinguished by the source of input energy:[8] (1) photoluminescence involving excitation by optical radiation, (2) cathodoluminescence by electron beam or cathode ray, (3) radioluminescence by other fast particles or high-energy radiation, and (4) electroluminescence by electric field or current. We are mainly concerned with electroluminescence here, especially with injection electroluminescence, that is, optical radiation obtained by injecting minority carriers into the vicinity of a semiconductor *p-n* junction where radiative transitions take place.

Electroluminescence may be excited in a variety of ways, including injection, intrinsic excitation, avalanche, and tunneling processes. Injection electroluminescence is by far the most-important method of excitation.[9] When a forward bias is applied to a *p-n* junction, the injection of minority carriers across the junction can give rise to efficient radiative recombination, since electric energy can be converted directly into photons. In subsequent sections we shall be concerned primarily with injection electroluminescent devices, that is, LEDs and lasers.



(a) (b)

Fig. 6 (a) GaAs diode emission spectra at 300 and 77 K. (b) Dependence of emission peak and half-power width as a function of temperature. (After Ref. 7.)

For the intrinsic excitation, a powder of a semiconductor (e.g., ZnS) is embedded in a dielectric (plastic or glass) and subjected to an alternating electric field. At frequencies in the audio range, electroluminescence usually occurs. Generally the efficiency is low (≤ 1%). The mechanism is mainly caused by impact ionization of accelerated electrons or field emission of electrons from trapping centers.[3,10]

For avalanche excitation, a *p-n* junction or a metal-semiconductor barrier is reverse-biased into avalanche breakdown. The electron-hole pairs generated by impact ionization may result in emission of either interband (avalanche emission) or intraband (deceleration emission) transitions. Electroluminescence can also result from tunneling in forward-biased or reverse-biased junctions. When a sufficiently large reverse bias is applied to a metal-semiconductor barrier (on *p*-type degenerate substrate) for example, holes at the metal side can tunnel into the valence band of the semiconductor and subsequently make a radiative recombination with electrons that have tunneled in opposite direction from the valence band to the conduction band.[11]

12.3 LIGHT-EMITTING DIODE (LED)

The light-emitting diode, commonly known as LED, is a semiconductor *p-n* junction that under proper forward-biased conditions can emit external spontaneous radiation in the ultraviolet, visible, and infrared regions of the spectrum. Electro-luminescence was first discovered by Round as early as 1907 in a contact to a SiC substrate, but was reported only in a short note.[1] More-detailed experiments were presented by Lossev, whose work spanned from the 1920s to the 1930s.[12,13] After the development of the *p-n* junction in 1949, LED structures changed from point contacts to *p-n* junctions. Other semiconductor materials besides SiC were subsequently studied, such as Ge and Si.[14] Since these semiconductors have indirect energy gap, their efficiencies were very limited. Much-higher quantum efficiencies were reported from direct-bandgap GaAs in 1962.[15–17] These studies quickly led to the realization of the semiconductor laser later the same year. Up to that point, it was considered imperative to use direct-bandgap material for efficient electroluminescence. Significant advancement was made on indirect-bandgap materials during 1964–1965 by introducing isoelectronic impurities.[18 20] These studies had a profound impact on commercial LEDs made from indirect-bandgap GaAsP and GaP. Most recently, a major advancement was made when InGaN was used to produce blue and UV portion of the spectrum which had not been possible before.[21] This technological advancement not only improves greatly on the realization and performance of white-light LEDs, it also helps to lift the popularity of LEDs as a whole.

The applications of LEDs are very wide and can be categorized into three kinds. The first is for display. Typical day-to-day examples are panel displays in different electronic equipments for audio and video home entertainment, panel displays in automobiles, computer screens, calculators, clocks and watches. Outdoor signs and traffic lights are getting increasingly popular as the efficiency and intensity keep on improving. Chances are everybody looks at some LED display everyday. Figure 7

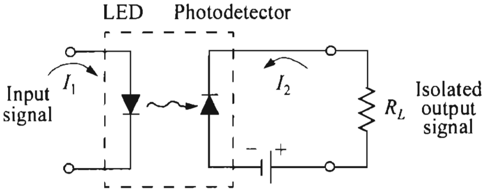


Fig. 8 An opto-isolator provides electrical isolation between input and output.

shows some basic formats for LED displays. The 7-segment configuration is usually used to display numbers from 0 to 9. For alphanumeric displays (A to Z and 0 to 9), the 5×7 matrix is common. The arrays of LEDs can be made by monolithic processes similar to those used to make silicon integrated circuits, or by mounting individual LEDs onto a package for larger displays.

The second category is for illumination, replacing the traditional incandescent light bulbs. Examples in this category are household lamps, flashlight, book light, automobile headlights, etc. The big advantage here is their high efficiencies, extending the battery life many times in portable usage. In addition, LEDs are more reliable and have longer lifetime. This feature greatly reduces the cost of having to replace traditional light bulbs, especially important in outdoor applications such as traffic lights.

The third application is as light source for optical-fiber communication systems, for low and medium data rates (< 1 Gb/s), over short and medium distances (< 10 km). Infrared LEDs are more suitable for this application since the wavelength ensures minimum loss in typical optical fibers. There are advantages and disadvantages to using LEDs as an optical source compared to semiconductor lasers. The advantages of LEDs include higher-temperature operation, smaller temperature dependence of emitted power, simpler device construction, and simpler drive circuit. The disadvantages are lower brightness and lower modulation frequency, and wide spectral line width, typically 5 to 20 nm as compared to the narrow line width, 0.1 to 1 Å, of a laser.

Similarly, LEDs can be used in opto-isolators where an input signal or control signal is decoupled from the output.[2] Figure 8 shows an opto-isolator having an LED

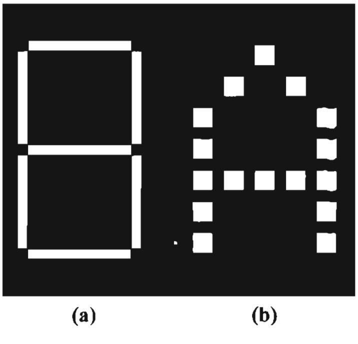

(a) (b)

Fig. 7 Typical configurations for (a) numerical display (7-segment) and (b) alphabetical display (5×7 array).

as the light source and coupled to a photodetector. When an input electrical signal is applied to the LED, light is generated and subsequently detected by the detector. The light is then converted back to an electrical signal as a current that can flow through a load resistor R_L. These devices are optically coupled with the signal transmitted at the speed of light, and are electrically isolated because there is no feedback or interference between the output and the input signals. In essence, this drawing also represents a optical-fiber communication system when the light is guided through a long-distance optical fiber.

12.3.1 Device Structures

The basic structure of an LED is a *p-n* junction. Under forward bias, minority carriers are injected from both sides of the junction. At the vicinity of the junction, there is an excess of carriers over their equilibrium values ($pn > n_i^2$), and recombination will take place. This condition is shown in Fig. 9a. However, if a heterojunction is utilized in the design, the efficiency can be much improved. Figures 9b shows that the central material where light is produced is bound by layers with a higher energy gap. If the heterojunctions are of Type-I (see Section 1.7), excess carriers of both types are injected and confined at the same space. As can be seen, the number of excess carriers can be significantly increased. It will be shown later that with increased carrier concentrations, the radiative recombination lifetime is shortened, leading to more-efficient radiative recombination. In this configuration, the central layer is undoped,

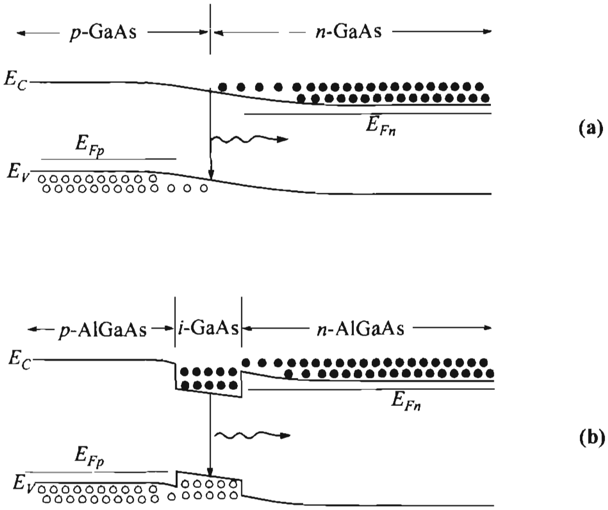


Fig. 9 (a) Under forward bias of a *p-n* junction, electrons injected from *n*-side recombine with holes injected from *p*-side. (b) Higher carrier densities and improved carrier confinement in a double heterojunction.

bound by layers of opposite types. This double-heterojunction design yields the highest efficiency and is the preferred approach.

Furthermore, if the central active layer is reduced to the range of 10 nm or smaller, a quantum well is formed. In this case, the 2-dimensional carrier densities become relevant and have to be calculated based on quantum mechanics. The effective 3-dimensional carrier densities (per unit volume), however, is given by the 2-D values divided by the quantum-layer width. This phenomenon pushes the carrier densities to higher levels and can result in higher efficiency. Another advantage of a thin active layer comes about in epitaxial growth, since a thin strained layer can accommodate higher level of lattice mismatch (see Section 1.7). Another feature of a quantum well is that quantization levels can theoretically extend the radiation energy (or to shorter wavelength) beyond the energy gap, but this feature is rarely used.

12.3.2 Materials of Choice

The most-important semiconductors for LED applications are listed in Figure 10, where the relative luminosity for the human eye is also added for reference. The spectrum covers most of the visible and extends into the infrared region. For display applications, since the human eye is only sensitive to light of energy $h\nu \geq 1.8$ eV ($\lambda \leq 0.7$ µm), semiconductors of interest must have energy bandgaps larger than this value. In general, all of these semiconductors are direct-bandgap materials except for some of the alloy composition in the GaAsP system which will be discussed in more details later. Direct-bandgap semiconductors are particularly important for electroluminescent devices, because the radiative recombination is a first-order transition process (no phonon involved) and the quantum efficiency is expected to be much higher than that for an indirect-bandgap semiconductor, where a phonon is involved.

AlGaAs. The $Al_xGa_{1-x}As$ systems covers a wide range of wavelengths from red to infrared. In the form of GaAs, it is the earliest material used for high-efficiency LEDs

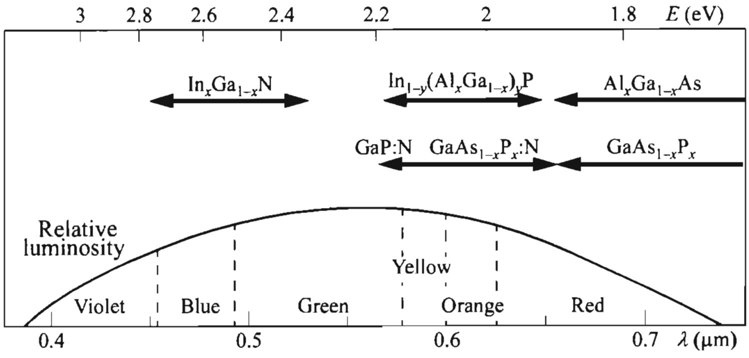


Fig. 10 Semiconductors of interest for LEDs, including the relative luminosity function of the human eye.

in the 1960s. For higher Al composition, ≈ 45%, its bandgap becomes indirect, so the wavelength is limited to about 0.65 μm. The advantages of this material system include excellent heterojunction growth capability for making double-heterojunction LEDs, and it has a good lattice match to GaAs substrates. GaAs has the most-advanced material technology of all the compound semiconductors.

InAlGaP. This material system has higher energy than AlGaAs, and it covers a wide range of the visible spectrum, i.e., red, orange, yellow, and green. The direct-bandgap range limits this material system to wavelengths longer than 0.56 μm. This system also has good lattice match to GaAs substrate.

InGaN. This most-recent technological breakthrough in InGaN epitaxial growth gives an important addition to LED applications. This material has a wide spectrum covering green, blue, and violet. More importantly, it is the sole provider for blue and violet which had been difficult from a material point of view. For longer wavelengths extending to the rest of the visible spectrum, higher In percentage is needed to decrease the energy gap. But higher In percentage is accompanied by more misfit dislocations because of increased lattice mismatch. Therefore, this material is not used for the rest of the visible spectrum. The substrate materials can be sapphire, SiC, or GaN, but the high cost of the latter two substrates encourages the use of sapphire.

GaAsP. As shown in Fig. 10, the $GaAs_{1-x}P_x$ system covers a very wide range of spectrum, from the infrared to the middle of the visible spectrum. The direct-indirect bandgap transition occurs at about 1.9 eV (P between 45–50%), and for wavelength generation in the indirect-bandgap regime, the efficiency is very low. However, an efficient radiative recombination center can be introduced by incorporating specific impurities such as nitrogen.[22]

When nitrogen is introduced, it replaces some of the phosphorus atoms in the lattice sites. The outer electronic structure of nitrogen is similar to that of phosphorus (both are Group-V elements in the periodic table), but the electronic core structures of these atoms differ considerably. This difference gives rise to an electron trap level close to the conduction band. A recombination center produced this way is called an *isoelectronic center.* ZnO is another type of isoelectronic center for GaP. The isoelectronic centers are normally neutral. Under operation, an injected electron is first trapped at this center. The negatively charged center then captures a hole from the valence band to form a bound exciton. The subsequent annihilation of this electron-hole pair yields a photon with an energy equal to the bandgap minus an energy approximately equal to the binding energy of the center. Such a system and operation can be visualized in the E-k diagram shown in Fig. 11. Conservation of momentum is not violated here because the isoelectronic traps are highly localized in space, and because of the Uncertainty principle, they have a wide range in the k-space (momentum).

Figure 12a shows the quantum efficiency versus alloy composition for $GaAs_{1-x}P_x$ with and without the isoelectronic impurity nitrogen.[22] The efficiency without nitrogen drops sharply in the composition range $0.4 < x < 0.5$ because of the proximity of the direct–indirect-E_g transition. The efficiency with nitrogen is considerably

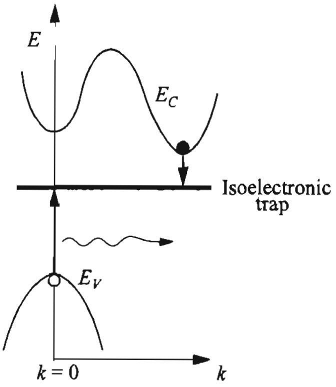


Fig. 11 *E-k* diagram showing radiative recombination through isoelectronic trap in indirect-bandgap material.

higher for $x > 0.5$, but still decreases steadily with increasing x because of the increasing separation in momentum between the direct and indirect bandgap (Fig. 4b). The nitrogen-doped alloy also shows a shift of the peak emission wavelength because of the binding energy of the isoelectronic center (Fig. 12b).

GaAsP LEDs can be grown on GaAs or GaP substrates, depending on the P content level. The advantage of the GaP substrate is its higher energy gap such that reabsorption of light by the substrate is minimal. LEDs with isoelectronic centers also have similar advantages because light emitted through these centers has reduced energy and thus is transparent to the substrate.

Wavelength Converters. A usual technique to tune the LED color is to cover the LED with a coating of a wavelength converter. This converter absorbs the LED light

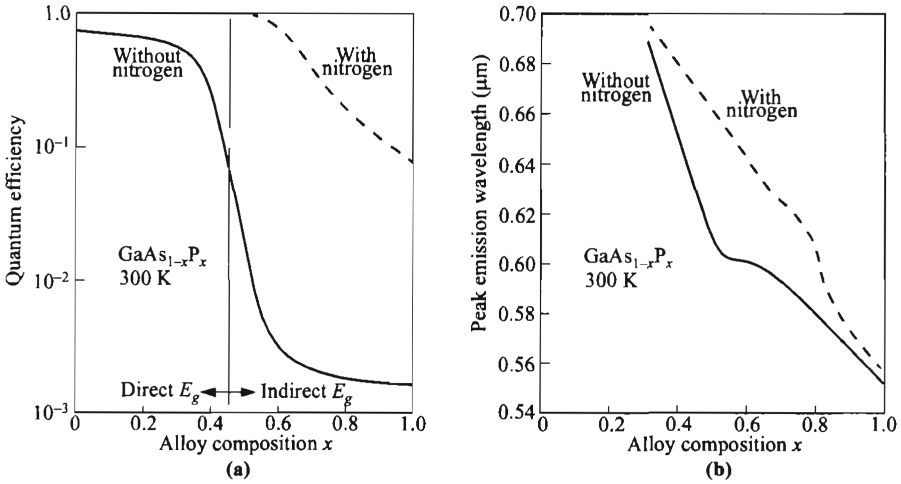


Fig. 12 (a) Quantum efficiency of $GaAs_{1-x}P_x$ vs. alloy composition, with and without isoelectronic impurity nitrogen. (b) Peak emission wavelength vs. composition. (After Ref. 22.)

and reemits light of different wavelengths. The wavelength converters can be phosphors,[23] dyes, and other semiconductors. They usually convert to lower energy (longer wavelength) and have a wider spectrum compared to the original light. Their efficiencies are usually quite high. These are the reasons that they are quite common in white-light LEDs. Not as common is conversion to higher energy. Blue LEDs can be obtained from infrared-to-visible up-converters,[24] where the infrared emission from a GaAs LED is absorbed by phosphor doped with rare-earth ions, such as ytterbium (Yb^{3+}) and erbium (Er^{3+}). The operation depends on the successive absorption of two photons in the infrared region followed by the emission of a single photon in the visible region.

12.3.3 Definitions of Efficiencies

The main function of an LED is to convert electrical energy into light in the visible part of the spectrum for display and illumination purposes. In this section, the different terms of LED efficiencies are discussed. Knowing the origins of these efficiencies, optimization can be made accordingly.

Internal Quantum Efficiency. For a given input power, the radiative recombination processes are in direct competition with the nonradiative ones. Each of the band-to-band transitions and transitions via traps can be either radiative or nonradiative. Examples of nonradiative band-to-band recombination are those in indirect-bandgap semiconductors. Conversely, examples of radiative recombination via traps are those via isoelectronic levels.

The internal quantum efficiency η_{in} is the efficiency of converting carrier current to photons, defined as

$$\eta_{in} = \frac{\text{number of photons emitted internally}}{\text{number of carriers passing junction}} . \tag{8}$$

It can be related to the fraction of the injected carriers that combine radiatively to the total recombination rate, and may also be written in terms of their lifetimes as

$$\eta_{in} = \frac{R_r}{R_r + R_{nr}} = \frac{\tau_{nr}}{\tau_{nr} + \tau_r} \tag{9}$$

where R_r and R_{nr} are the radiative and nonradiative recombination rates, and τ_r and τ_{nr} are their associated radiative and nonradiative lifetimes, respectively. For low-level injection, the radiative recombination rate in the p-side of the junction is given by

$$\begin{aligned} R_r &= R_{ec} np \\ &\approx R_{ec} \Delta n N_A , \end{aligned} \tag{10}$$

where R_{ec} is the recombination coefficient and Δn is the excess carrier density which is much larger than the minority carrier density in equilibrium $\Delta n \gg n_{po}$. R_{ec} is a function of the band structure and temperature. Its value would be very small for indirect-bandgap semiconductors. ($R_{ec} \approx 10^{-10}$ cm^3/s for direct-bandgap materials, and $\approx 10^{-15}$ cm^3/s for indirect-bandgap materials.)

For low-level injection ($\Delta n < p_{po}$), the radiative lifetime τ_r is related to the recombination coefficient by

$$\tau_r = \frac{\Delta n}{R_r} = \frac{1}{R_{ec}N_A}. \tag{11}$$

For high-level injection, however, τ_r would decrease with Δn. So in double-heterostructure LEDs, carrier confinement increases Δn and τ_r is reduced to improve the internal quantum efficiency. The nonradiative lifetime is usually attributed to traps (of density N_t) or recombination centers,

$$\tau_{nr} = \frac{1}{\sigma v_{th} N_t} \tag{12}$$

where σ is the capture cross section. It is evident that the radiative lifetime τ_r needs to be small to yield high internal quantum efficiency.

External Quantum Efficiency. Obviously for LED applications, what matters is the light emitted external to the device. For this, the optics inside and outside the device has to be considered. The parameter to measure the efficiency of getting the light out externally is the *optical efficiency* η_{op}, sometimes called the *extraction efficiency.* With this factored in, the net *external quantum efficiency* is defined as

$$\eta_{ex} = \frac{\text{number of photons emitted externally}}{\text{number of carriers passing junction}} = \eta_{in}\eta_{op}. \tag{13}$$

The optical efficiency is a subject of optics inside and around the devices, totally independent of electrical phenomena. We focus on the device optical paths and optical interfaces in the following section.

Optical Efficiency. First we present the basic law of refraction when light passes through the semiconductor-ambient interface, shown in Fig. 13. Most of the important phenomena arise from Snell's law, which states that the directions of light before (θ_s) and after (θ_o) the interface is governed by

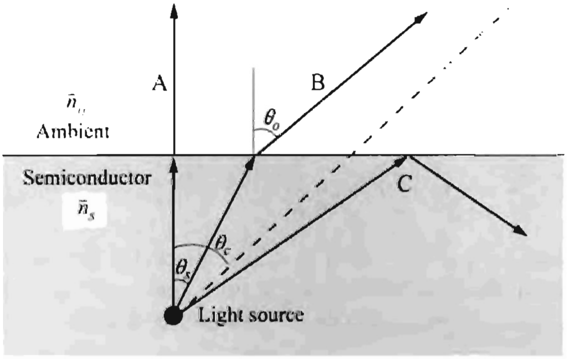


Fig. 13 Optical paths at the semiconductor/ambient interface. A: Normal incidence has little effect. B: Angles of refraction ($\theta_o > \theta_s$) corresponding to Snell's law. C: Ray outside the light-escape cone ($\theta_s > \theta_c$) has total reflection.

$$\bar{n}_s \sin\theta_s = \bar{n}_o \sin\theta_o , \tag{14}$$

where $\bar{n}_s$ and $\bar{n}_o$ are the refractive indexes of the semiconductor and ambient respectively. For normal incidence, the direction of path is not changed, except it suffers from the Fresnel loss with a reflection coefficient

$$R = \left(\frac{\bar{n}_s - \bar{n}_o}{\bar{n}_s + \bar{n}_o}\right)^2 . \tag{15}$$

For optical paths with $\theta_s > 0°$, since $\bar{n}_s$ (around 3–4 for common semiconductors) is larger than $\bar{n}_o$ (air = 1), θ_o is always larger than θ_s. Figure 13 shows that there is a critical angle θ_c for θ_s when θ_o becomes 90° and the refracted light is parallel to the interface. Such a critical angle defines the light-escape cone outside which light is totally reflected back to the semiconductor. Substituting $\theta_o = 90°$ in Eq. 14, this critical angle is given by

$$\theta_c = \sin^{-1}\left(\frac{\bar{n}_o}{\bar{n}_s}\right) \approx \frac{\bar{n}_o}{\bar{n}_s} . \tag{16}$$

For GaAs ($\bar{n}_s = 3.66$) and GaP ($\bar{n}_s = 3.45$), the critical angle is about 16°–17°.

Three major loss mechanisms reduce the quantity of emitted photons: (1) absorption within the LED material, (2) Fresnel loss, and (3) critical-angle loss. The absorption loss for LEDs on GaAs substrates is large since the substrate is opaque to light and it absorbs about 85% of the photons emitted at the junction. For LEDs on transparent substrates such as GaP with isoelectronic centers, photons emitted downward can be reflected back with only about 25% absorption; the efficiency can be significantly improved. The Fresnel loss is due to internal reflection back to the semiconductor. The third loss mechanism is caused by the total internal reflection of photons incident to the surface at angles greater than the critical angle θ_c.

To estimate the optical efficiency due to the critical-angle loss, we ignore the absorption loss and the Fresnel loss for the sake of simplicity. The solid angle of the light-escape cone can be calculated to be

$$\text{Solid angle} = 2\pi(1 - \cos\theta_c) . \tag{17}$$

While the total solid angle from a point source is 4π, the optical efficiency can be simply given by the fraction

$$\begin{aligned} \eta_{op} &= \frac{\text{solid angle of light-escape cone}}{4\pi} = \frac{1}{2}(1 - \cos\theta_c) \\ &= \frac{1}{2}\left[1 - \left(1 - \frac{\theta_c^2}{2!} + \ldots\right)\right] \approx \frac{1}{4}\theta_c^2 \approx \frac{1}{4}\frac{\bar{n}_o^2}{\bar{n}_s^2} . \end{aligned} \tag{18}$$

(Series expansion for $\cos\theta_c$ has been used.) For a typical semiconductor LED with planar surface, it is seen here that the optical efficiency is only in the order of 2%.

One interesting phenomenon arising from the Snell's law is that even though light inside the semiconductor has uniform intensity, light emitted into the ambient after refraction at the interface has an angle dependence. It has a maximum intensity when the light is normal to the interface, and decreases when the angle θ_o is increased.

Equating the light energy below and above the interface, it can be shown that for a common planar LED structure, the emitted light intensity has an angle dependence of

$$I_o(\theta_o) = \frac{P_s}{4\pi r^2}\frac{\bar{n}_o^2}{\bar{n}_s^2}\cos\theta_o, \tag{19}$$

where P_s is the power of the light source and r the distance of the surface from the source. Such an emission pattern is called the Lambertian emission pattern. Figure 14 shows such emission patterns for a planar structure, a hemispherical structure, and a parabolic structure. It can be seen that for a planar structure, at an angle of 60° the normalized intensity drops to 50%. For an ideal hemisphere, since all rays are normal to the interface, the emitted intensity maintains a uniformly high intensity and the critical-angle loss is totally eliminated. However, in practice such a hemispherical shape is very hard to achieve. A good practical compromise is to cap the planar structure with a hemispherical coating whose refractive index lies between that of the semiconductor and the ambient.

The total emitted optical energy of a planar structure can be calculated by integrating Eq. 19 over the entire rang of 0° ≤ θ_o ≤ 90°. The optical efficiency can be calculated by comparing the emitted light power to the power source at the junction. The optical efficiency calculated in this manner will give the same end result as Eq. 18.

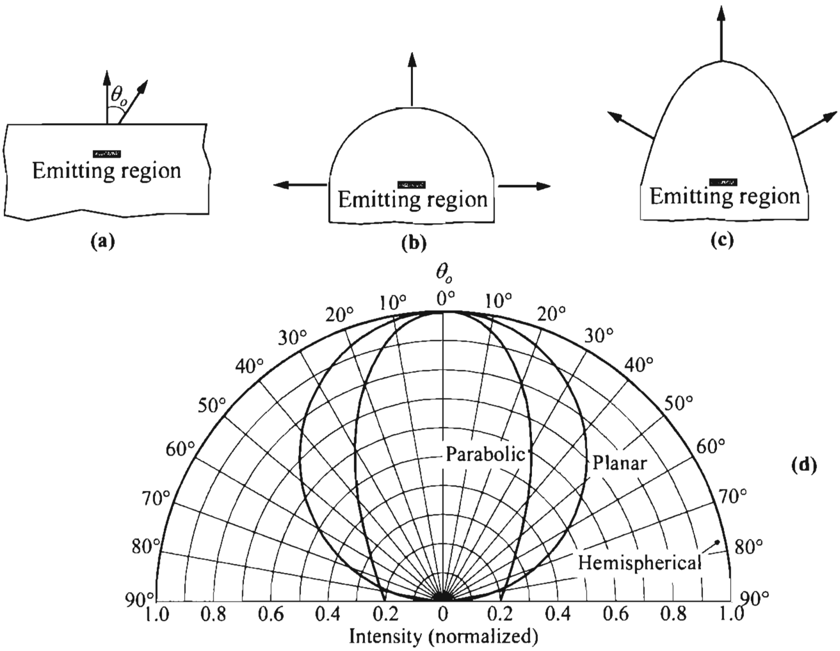

θ_o
Emitting region
(a)
Emitting region
(b)
Emitting region
(c)
θ_o
0°
10°
10°
20°
20°
30°
30°
40°
40°
50°
50°
60°
60°
70°
70°
80°
80°
90°
90°
Parabolic
Planar
Hemispherical
1.0
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1.0
Intensity (normalized)
(d)

Fig. 14 LED structures for optical-efficiency consideration: (a) Planar. (b) hemispherical. (c) Parabolic. (d) Their normalized Lambertian emission patterns. (After Ref. 6.)

So far, we have considered that light is emitted out of the device in the direction away from the junction and out of the top or bottom surface. Such devices are called surface emitters. Another device option, called edge emitters, has light coming out parallel to the junction (Fig. 15). These are the two basic device configurations to couple the LED light output into a small glass fiber. For the surface emitter (Fig. 15a), the emitting area of the junction is confined by oxide isolation and a least-resistive path formed by the p^+-diffusion. The semiconductor layer that the emission must pass through is made very thin, 10 to 15 μm, to minimize absorption. The use of heterojunctions (e.g., GaAs/AlGaAs) can increase the efficiency resulting from carrier confinement provided by the layers of higher bandgap semiconductor (e.g., AlGaAs) surrounding the radiative recombination region (e.g., GaAs). The heterojunction can also serve as a window to the emitted radiation, because the higher bandgap confining layers do not absorb radiation from the lower bandgap emitting region. For the edge emitters (Fig. 15b), the active layer and the double-heterostructure is sandwiched by two optical cladding layers, and, in effect, a waveguide is formed. Their light outputs are more collimated so they do not suffer from total reflection associated with the critical angle. They have the advantage of improved efficiency in coupling to a fiber with a small acceptance angle. The spatial distribution of the emitted light is similar to the distribution of a heterostructure laser, considered in Section 12.5.4.

Power Efficiency. The power efficiency η_P is simply defined as the ratio of the light power output to the electrical power input,

$$\eta_P = \frac{\text{optical power out}}{\text{electrical power in}} = \frac{\text{number of photons emitted externally} \times h\nu}{I \times V}$$
$$= \frac{\text{number of photons emitted externally} \times h\nu}{\text{number of carriers passing junction} \times q \times V} . \tag{20}$$

Since the bias is approximately equal to the energy gap and light energy ($qV \approx h\nu$), it follows that the power efficiency is similar to the external quantum efficiency ($\eta_P \approx \eta_{ex}$).

Luminous Efficiency. When comparing the visual effects of LEDs, the eye response must also be taken into account. The luminous efficiency normalizes the power effi-

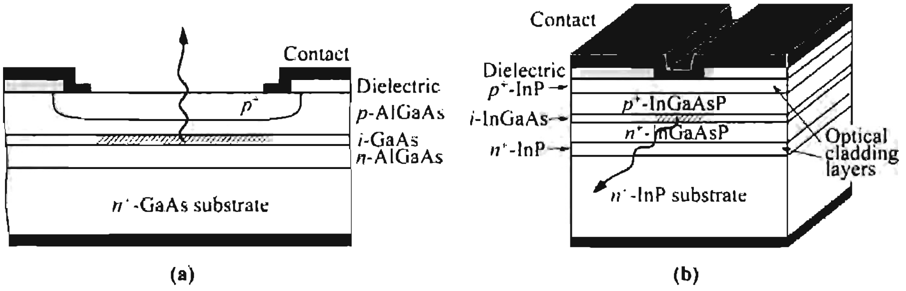


Fig. 15 LED structures showing the direction of emitted light from (a) surface emitter and (b) edge emitter.

ciency by a factor that is related to the eye sensitivity as shown in Figure 1 earlier. For example, the human eye has a peak sensitivity at 0.555 μm (green). As the wavelength approaches the red end or violet end of the visible spectrum, the sensitivity falls rapidly. So it takes less power in green color than those in other colors to achieve the same visual brightness. For LED applications in display and illumination, the luminous efficiency is a more appropriate parameter.

The brightness of light output is measured by the luminous flux (in lumens),

$$\text{luminous flux} = L_0 \int V(\lambda) P_{\text{op}}(\lambda) d\lambda \qquad \text{lm}, \tag{21}$$

where L_0 is a constant with a value of 683 lm/W, $V(\lambda)$ the relative eye sensitivity (Fig. 1), and $P_{\text{op}}(\lambda)$ the power spectrum of the radiation output. The eye sensitivity function $V(\lambda)$ is normalized to unity for the peak at $\lambda = 555$ nm. The luminous efficiency is then given by[9]

$$\eta_{lu} = \frac{\text{luminous flux}}{\text{electrical power in}} = \frac{683 \int V(\lambda) P_{\text{op}}(\lambda) d\lambda}{VI} \qquad \text{lm/W}. \tag{22}$$

The maximum luminous efficiency has a value of 683 lm/W.

As LED technology advances with time, the luminous efficiency has achieved impressive progress. The chronological improvement of the luminous efficiency is summarized in Fig. 16. The luminous efficiencies for conventional lightings are also included for comparison. The slope in the figure shows an improvement of a factor of two for every 3 years, or equivalently tenfold per decade. Obviously such rate of improvement cannot be sustained as the luminous efficiency approaches the theoretical limit of 683 lm/W. By now, the most-advanced LEDs have luminous efficiencies already surpassing those of the traditional lightings.

12.3.4 White-Light LED

One important application is to use white-light LEDs for general-purpose high-brightness illumination.[25] This area of application is becoming more and more important as both the power efficiency and the brightness have been improved to the extent that it is in direct competition with conventional lightings, i.e., incandescent and fluorescent lightings. Examples of day-to-day usage are house lamps, decorative lights, flashlights, outdoor signs, automobile headlights, etc.

White light can be produced by mixing two or three colors of an appropriate intensity ratio. There are basically two approaches to achieve while light. The first is to combine LEDs of different colors: red, green, and blue. This is not a popular approach since it is more costly, and mixing of multiple colors of narrow bandwidth does not produce good color rendering. The second approach, most commonly used, is to have a single LED covered with a color converter. A color converter is a material that absorbs the original LED light and emits light of different frequency. The converter material can be phosphor, organic dye, or another semiconductor. Of the three materials, phosphor is the most common.[23] The light output from a phosphor generally has

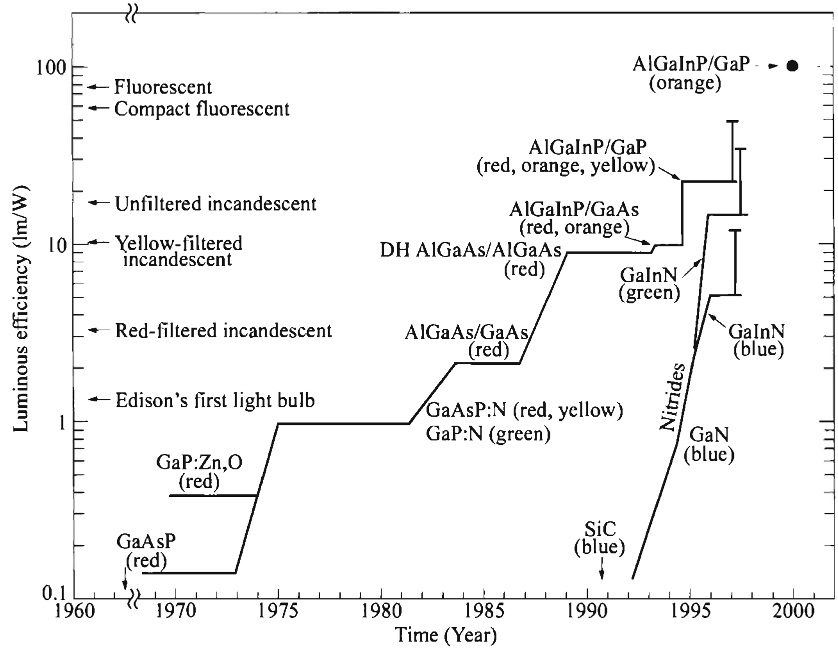


Fig. 16 Progression of LED luminous efficiency with time. (After Ref. 6.)

a much broader spectrum compared to the LED light, and the wavelength range is longer (lower photon energy). The efficiencies of these color converters can be very high, near 100%.

One popular version is to use a blue LED together with a yellow phosphor. In this scheme, the LED light is partially absorbed by phosphor. The blue LED light is mixed with yellow light produced by the phosphor to give white light. Another version is to use a UV LED. The LED light is completely absorbed by the phosphor, and a wide spectrum of light is reproduced that emulates white light.

12.3.5 Frequency Response

The frequency response is another important parameter to be considered in the design of LEDs for high-speed applications such as optical-fiber communication systems. It determines the maximum frequency at which the LEDs can be turned on and off, and, thus, the maximum transmission rate of data. This cutoff frequency of an LED is given by

$$f_T = \frac{1}{2\pi\tau}, \tag{23}$$

where τ is the overall lifetime defined as

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}}. \tag{24}$$

As discussed earlier, the internal quantum efficiency is related to the radiative and nonradiative lifetimes, τ_r and τ_{nr}. In Eq. 24, τ approaches τ_r when $\tau_r \ll \tau_{nr}$. Therefore, as Eq. 11 indicates, τ_r decreases as the doping in the active layer is increased, and f_T becomes larger. So for speed consideration, increased doping concentration in the middle active layer of the heterostructure is desirable.[26]

12.4 LASER PHYSICS

Laser (light amplification by stimulated emission of radiation) is the descendant of *maser* (microwave amplification by stimulated emission of radiation). The difference between them is in the range of output frequencies. The laser and maser are both based on the phenomenon of stimulated emission, which was postulated by Einstein in the 1910s. The laser medium can be gas, liquid, amorphous solid, or semiconductor. The semiconductor laser is also called injection laser, junction laser, or laser diode.

Maser action was first realized by Townes and his collaborators[27] and by Basov and Prokhorov,[28] both in 1954, using ammonia gas. Laser action was obtained first on solid ruby (nonsemiconductor) in 1960[29] and then on helium-neon gas in 1961. Subsequently semiconductors were suggested for use as laser materials.[30–32] The theoretical calculations of Bernard and Duraffourg[33] in 1961 and Dumke[34] in 1962 showed that laser action was indeed possible in direct-bandgap semiconductors and set forth important criteria for such action. In 1962, four papers reported the semiconductor lasers almost simultaneously: Hall et al.,[35] Nathan et al.,[36] and Quist et al.[37] on GaAs, and Holonyak and Bevacqua on GaAsP.[38] Improvement based on heterojunctions was suggested by Kroemer[39] and by Alferov and Kazarinov[40] in 1963. Eventually, Hayashi et al. in 1970 achieved CW operation at room temperature using a double-heterojunction laser.[41] The historical development of the laser, from maser to heterojunction laser, may be found collectively in Refs. 42–44.

Semiconductor lasers are similar to other lasers (such as the solid-state ruby laser and He-Ne gas laser) in that the emitted radiation has spatial and temporal coherence. Laser radiation is highly monochromatic (of small bandwidth) and it produces highly directional beams of light. However, semiconductor lasers differ from other lasers in several important respects:

1. In conventional lasers, the quantum transitions occur between discrete atomic energy levels, whereas in semiconductor lasers the transitions are associated with the band properties of materials.
2. A semiconductor laser is very compact in size, on the order of 0.1-mm long for discrete lasers. More recently, integrated lasers made on the monolithic wafer form are much smaller. However, because the active region is very narrow (on

the order of 1-μm thick or less), the divergence of the laser beam is considerably larger than in a conventional laser.

3. The spatial and spectral characteristics of a semiconductor laser are strongly influenced by the properties of the junction medium such as bandgap and refractive index variations.
4. The laser action is pumped simply by passing a forward current through the diode, as opposed to optical pumping for example. The result is a very efficient overall system that can be modulated easily by modulating the current. Since semiconductor lasers have very short photon lifetimes, modulation at high frequencies can be achieved.

Since the initial discoveries, many semiconductor materials have been found to lase in coherent radiation extending from the near ultraviolet into the visible and out to the far-infrared spectrum (wavelength ≈ 0.2 to ≈ 40 μm). Semiconductor lasers, because of their wavelength tunability along with narrow spectral linewidth, high stability, low input power, and structural simplicity, have significant potential for application in technology and basic research, such as molecular spectroscopy, atomic spectroscopy, high-resolution gas spectroscopy, and monitoring atmospheric pollution. The applications of semiconductor lasers cover an extremely wide range from many areas of basic research to medical surgeries to day-to-day consumer electronics. Because of its compact size and capability for high-frequency modulation, the semiconductor laser is the most-important light source for the optical-fiber communication system. At the same time, recent technological development has brought the cost low enough so that the most-popular applications are in the consumer market of CD and DVD players.

12.4.1 Stimulated Emission and Population Inversion

To get a clear picture, we start with a simpler atomic system having two discrete energies, as opposed to bands of energy in semiconductors. Consider two energy levels E_1 and E_2, where E_1 is the ground state and E_2 is an excited state (Fig. 3). These energy states have electron concentrations of N_1 and N_2. Any transition between these states involves the emission or absorption of a photon with frequency ν given by $h\nu = E_2 - E_1$. As mentioned previously, the three optical processes are absorption, spontaneous emission, and stimulated emission (of transition rates R_{ab}, R_{sp}, and R_{st} respectively). At ordinary temperatures, most of the atoms are in the ground state. This situation is disturbed when a photon of energy $h\nu$ impinges on the system. An atom in state E_1 absorbs the energy and thereby goes to the excited state E_2. This is the absorption process. Absorption is characterized by the absorption coefficient (α) and is the principal process in photodetectors and solar cells. The excited state of the atom is unstable and after a short time, without any external stimulus, makes a transition to the ground sate giving off a photon of energy $h\nu$. This process is called spontaneous emission. The lifetime for spontaneous emission (i.e., the lifetime of the excited state) varies considerably, ranging typically from 10^{-9} to 10^{-3} s, depending on various semiconductor parameters such as bandgap (direct or indirect) and density of recombination centers. In spontaneous emission, light produced is random in space and time (incoherent). It is the dominant mechanism in an LED. An important and interesting event occurs when a photon of energy $h\nu$ impinges on an atom while it is still in the excited state. In this case, the atom is immediately stimulated to make its transition to the ground state and gives off another photon of the same wavelength and is in phase with the incident radiation. This process is called stimulated emission. Stimulated emission is the main mechanism for lasing. Note the two interesting properties of the stimulated emission. First, one photon input is needed and it becomes two photons in the output, a basic concept of optical gain. Second, the two photons are in phase, making the laser output coherent.

We next proceed to analyze the basic requirement for stimulated emission. The formulae for the transition rates of the three optical processes are:

$$R_{ab} = B_{12} N_1 \phi , \tag{25}$$

$$R_{sp} = A_{21} N_2 , \tag{26}$$

$$R_{st} = B_{21} N_2 \phi . \tag{27}$$

B_{12}, A_{21}, B_{21} are called the Einstein coefficients for stimulated absorption, spontaneous emission, and stimulated emission respectively. Note that both R_{ab} and R_{st} are proportional to the light intensity ϕ, while R_{sp} is independent of the latter. In equilibrium, the ratio of electron concentrations residing in these states are related to their energies and is given by the Boltzmann statistics:

$$\frac{N_2}{N_1} = \exp\left(\frac{-\Delta E}{kT}\right) = \exp\left(\frac{-h\nu}{kT}\right) . \tag{28}$$

Recall that the black-body radiation has the intensity spectrum of

$$\phi(\nu) = \frac{8\pi \bar{n}_r^3 h\nu^3}{c^3}\left[\frac{1}{\exp(h\nu/kT) - 1}\right] . \tag{29}$$

Since the net optical transition is zero, we set $R_{ab} = R_{sp} + R_{st}$, giving

$$B_{12} N_1 \phi = N_2 (A_{21} + B_{21}\phi) . \tag{30}$$

Substituting Eqs. 28 and 29 into Eq. 30 gives the following general relationship

$$\frac{8\pi \bar{n}_r^3 h\nu^3}{c^3[\exp(h\nu/kT) - 1]} = \frac{A_{21}}{B_{12}\exp(h\nu/kT) - B_{21}} . \tag{31}$$

In order that this holds for all temperatures, the conclusion is that

$$B_{12} = B_{21} . \tag{32}$$

And it follows that

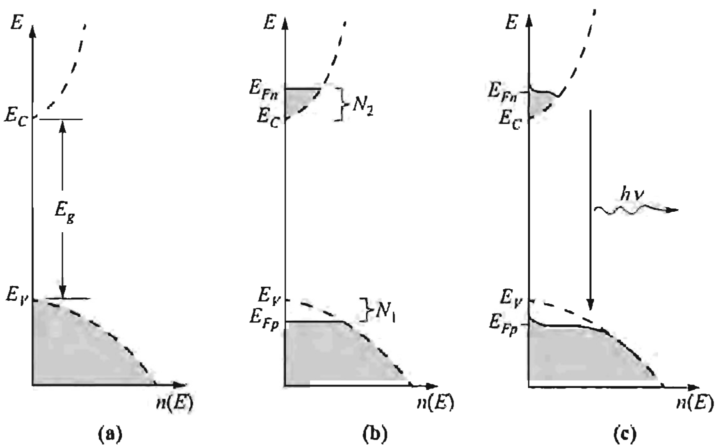
$$\frac{A_{21}}{B_{21}} = \frac{8\pi \bar{n}_r^3 h\nu^3}{c^3} . \tag{33}$$

In the laser action, the spontaneous emission for incoherent light is unimportant and weak, and it can be ignored. The net optical output is stimulated emission minus the absorption:

$$R_{st} - R_{ab} = (N_2 - N_1)B_{21}\phi. \tag{34}$$

It is seen here that the net optical gain is positive only if $N_2 > N_1$, a condition called *population inversion*. Under thermal equilibrium, according to Eq. 28, more atoms are in the ground states than in the excited state and population inversion does not occur naturally. Some external means is needed to create this state of population inversion. It can be provided by another optical source (optical pumping), or in the case of a laser diode, by forward biasing the *p-n* junction which is the basic device structure of a semiconductor laser.

Now we switch to semiconductors whose energy levels now become two separate continuous bands. To be consistent with the picture of population inversion, the concept of holes in the valence band is put aside for now. The region near the metallurgical junction where light is produced is shown in Fig. 17. At $T = 0$ and under equilibrium (Fig. 17a), states in the conduction band are completely empty of electrons and those in the valence band are completely full. Figure 17b shows the situation with population inversion at 0 K. This nonequilibrium condition can be characterized by two quasi-Fermi levels E_{Fn} and E_{Fp}. The conduction band is filled with electrons up to E_{Fn}, and the valence band empty of electrons down to E_{Fp}. Each level of E_1 and E_2 is now broadened into some narrow bands of interest: namely ($E_C \rightarrow E_{Fn}$) and ($E_{Fp} \rightarrow E_V$). N_1 and N_2 are integrated electron densities within these narrow bands. So in the example shown, N_2 is the total number of electrons in the conduction band, but N_1 is the number within the narrow band of ($E_{Fp} \rightarrow E_V$) and is zero. At a finite temperatures $T > 0$ K, the carrier distributions will be *smeared* out in energy, and the distribution function is no longer a step (Fig. 17c). Although overall thermal equilibrium does not exist, the electrons in a given energy band will be in thermal equilibrium



(a) (b) (c)

Fig. 17 Electron concentration as a function of energy in a semiconductor, determined by density of states (dashed) and Fermi-Dirac distribution. (a) Equilibrium, $T = 0$ K. With population inversion, (b) $T = 0$ K and (c) $T > 0$ K.

with each other. The occupation probability of a state in the conduction band and valence band is given by the Fermi-Dirac distribution

$$F_C(E) = \frac{1}{1+\exp[(E-E_{Fn})/kT]}, \tag{35a}$$

$$F_V(E) = \frac{1}{1+\exp[(E-E_{Fp})/kT]}. \tag{35b}$$

Consider the rate of photon emission at $h\nu$ due to a transition from upper states at E in the conduction band to lower states at $(E - h\nu)$ in the valence band. The rate for this emission is proportional to the product of the density of occupied upper states $F_C N_c$ and the density of unoccupied lower states $(1 - F_V)N_v$, where N_c and N_v are densities of states in the conduction band and valence band respectively. On the other hand, the rate of absorption is proportional to the product of density of unoccupied upper states $(1 - F_C)N_c$ and the density of occupied lower states $F_V N_v$. The transition rates for absorption R_{ab}, spontaneous emission R_{sp}, and stimulated emission R_{st} are thus given by integrating all energies

$$R_{ab} = B_{12}\int (1-F_C)F_V N_c N_v N_{ph} dE, \tag{36}$$

$$R_{sp} = A_{21}\int F_C(1-F_V)N_c N_v dE, \tag{37}$$

$$R_{st} = B_{21}\int F_C(1-F_V)N_c N_v N_{ph} dE. \tag{38}$$

N_{ph} is the density of photons of appropriate energy. For lasing consideration, the spontaneous emission is again ignored, and the net optical gain is given by

$$R_{st} - R_{ab} = B_{21}\int N_{ph}(F_C - F_V)N_c N_v dE \tag{39}$$

(using the equality $B_{12} = B_{21}$ derived earlier). In order for Eq. 39 to be positive, $F_C > F_V$, and $E_{Fn} > E_{Fp}$, which is the condition of population inversion for semiconductors. Recall that for thermal equilibrium, $E_{Fn} = E_{Fp}$ and $pn = n_i^2$, so our conventional way of representing the condition of population inversion can be simply $pn > n_i^2$.

Furthermore, from Eqs. 35a and 35b ($\Delta E = h\nu$), and with the requirement that the photon energy has to be larger than the bandgap, $h\nu > E_g$, the necessary condition for lasing becomes[33]

$$E_g < h\nu < E_{Fn} - E_{Fp}. \tag{40}$$

A qualitative picture on this requirement can be gained from Fig. 17c where the photon energy range is indicated.

Equation 40 also has important implication on the doping levels of the *p-n* junction. For a regular current-pumped laser diode, the quantity $(E_{Fn} - E_{Fp})$ is simply equal to the bias voltage. Since the bias is limited to the built-in potential of the junction, which is given by $(\psi_{Bn} + \psi_{Bp})$, it follows that the following requirement has to be met,

$$E_g < (\psi_{Bn} + \psi_{Bp})q\,. \tag{41}$$

This means at least one side of the junction (for homojunction) has to be highly doped to degeneracy such that its Fermi level lies within the band (or bulk potential ψ_B larger than half of the bandgap). However, for a heterojunction laser (discussed later), E_g for the light-emitting area is smaller, and the doping requirement is relaxed.

12.4.2 Optical Resonator and Optical Gain

Another structural requirement for a laser is an optical resonator in the direction of the light output. The optical resonator mainly serves to trap the light and build up the intensity inside. In the form of a Fabry-Perot etalon, it has two perfectly parallel walls and are perpendicular to the junction. These walls have mirror-like smoothness with an optimum design for reflectivity. One of the Fabry-Perot mirrors can be totally reflecting so that light comes out from only one side. The mirror walls parallel to the laser output are roughened to be highly absorbing to prevent lasing in the transverse direction.

The optical resonator has multiple resonant frequencies called longitudinal modes. Each corresponds to a standing wave with zero nodes at the boundaries. Under this condition, the repeatedly reflected rays are in phase with the rest inside the cavity, and positive interference preserves the state of coherence. This condition is satisfied when L is a multiple of the half wavelength, or

$$m\left(\frac{\lambda}{2\bar{n}_r}\right) = L\,, \tag{42}$$

where m is an integer. The separations of these modes in wavelength and frequency are given by

$$\Delta\lambda = \frac{d\lambda}{dm}\Delta m = \frac{\lambda^2}{2L\bar{n}_r}\Delta m\,, \tag{43}$$

$$\Delta\nu = \frac{c}{2L\bar{n}_r}\Delta m\,. \tag{44}$$

A typical length L is much larger than the wavelength of interest, so a precise dimension is not required.

Within the optical resonator, the optical gain (g) due to stimulated emission is compensated by optical loss due to absorption (α). The net gain/loss as a function of distance is given by

$$\phi(z) \propto \exp[(g - \alpha)z]\,. \tag{45}$$

Considering a complete return path, with reflections of R_1 and R_2 of the two mirrors (Fig. 18) these contribute to additional losses. Since for a given system, R_1, R_2, and α are fixed, g is the only parameter to vary the over-all gain. In order to have the over-all gain to be positive, the criterion is given by

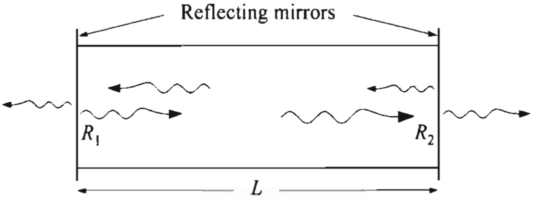

Reflecting mirrors
R_1
R_2
L

Fig. 18 Fabry-Perot optical cavity. R_1 and R_2 are reflection coefficients of the two mirrors.

$$R_1 R_2 \exp[(g-\alpha)2L] > 1 \quad . \tag{46a}$$

Or equivalently, the threshold gain g_{th} for lasing is

$$g_{th} = \alpha + \frac{1}{2L}\ln\left(\frac{1}{R_1 R_2}\right) . \tag{46b}$$

Since the gain is directly related to the pumping current, this criterion is the basis in determining the threshold current for lasing and is an important parameter.

12.4.3 Waveguiding

In the previous section, the optical resonator is shown to be critical in trapping the light and building up the intensity. This optical resonator is formed by mirrors perpendicular to the direction of the light. In this section, we discuss the confinement of light in the direction parallel to light propagation (so as to avoid leaking in the vertical direction in Fig. 18). This light confinement is provided by waveguiding, arising from nonuniform refractive indexes near the light-emitting junction. There is a large benefit provided in a double-heterojunction laser where the active layer is formed with material of higher refraction index than the surrounding layers, thereby forming a waveguide. Figure 19 shows a three-layer dielectric waveguide with refractive indexes, $\bar{n}_{r1}$, $\bar{n}_{r2}$, and $\bar{n}_{r3}$. Under the condition

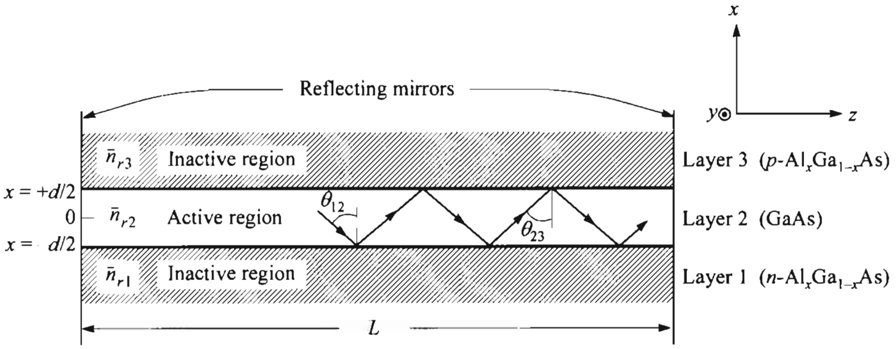

x
y
z
Reflecting mirrors
$\bar{n}_{r3}$
Inactive region
Layer 3 (p-Al$_x$Ga$_{1-x}$As)
$x = +d/2$
0
$\bar{n}_{r2}$
Active region
θ_{12}
θ_{23}
Layer 2 (GaAs)
$x = d/2$
$\bar{n}_{r1}$
Inactive region
Layer 1 (n-Al$_x$Ga$_{1-x}$As)
L

Fig. 19 Representation of a three-layer dielectric waveguide and ray trajectories of the guided wave.

$$\bar{n}_{r2} > \bar{n}_{r1}\ ,\ \bar{n}_{r3} \tag{47}$$

the ray angle θ_{12} at the layer-1/layer-2 interface in Fig. 19 exceeds the critical angle given by Eq. 16. A similar situation for θ_{23} occurs at the layer-2/layer-3 interface. Therefore, when the refractive index in the active region is larger than those of its surrounding layers, Eq. 47, the propagation of electromagnetic radiation is guided in a direction parallel to the layer interfaces.

For the homostructure laser, the difference in the refractive indexes between the center waveguiding layer and the adjacent layers is due to a different mechanism: material of higher carrier density has a lower refractive index. Here the active layer is more lightly doped and is sandwiched between heavily doped n^+- and p^+-layers. The difference in refractive indexes is only 0.1% to about 1%. For heterostructure lasers, the refractive index steps at each heterojunction can be made larger ($\approx$ 10%) and provide a well-defined waveguide.

To derive the detailed waveguiding properties rigorously, the transverse coordinates x and y coincide with the directions perpendicular and parallel to the junction plane, respectively. Consider a symmetric three-layer dielectric waveguide with $\bar{n}_{r2} > \bar{n}_{r1} = \bar{n}_{r3}$ (Fig. 19). For transverse electric (TE) waves polarized transversely to the direction of propagation (z-direction), $\mathscr{E}_z$ equals 0. The waveguide is considered to extend to infinity in the y-direction, so that $\partial/\partial y = 0$. The Maxwell wave equation is simplified to

$$\frac{\partial^2 \mathscr{E}_y}{\partial x^2} + \frac{\partial^2 \mathscr{E}_y}{\partial z^2} = \mu_0 \varepsilon \frac{\partial^2 \mathscr{E}_y}{\partial t^2} \tag{48}$$

where μ_0 is the permeability and ε is the dielectric permittivity. The solution, by separation of variables for even TE waves within the active layer $-d/2 < x < d/2$, is given by

$$\mathscr{E}_y(x, z, t) = A_e \cos(\kappa x) \exp[j(\omega t - \beta z)] \tag{49}$$

with

$$\kappa^2 \equiv \bar{n}_{r2}^2 k_0^2 - \beta^2 \tag{50}$$

where $k_0 \equiv (\omega/\bar{n}_{r2})\sqrt{\mu_0 \varepsilon}$ and β is the separation constant. The magnetic field in the z-direction is given by

$$\begin{aligned} \mathscr{H}_z(x, z, t) &= \left(\frac{j}{\omega \mu_0}\right) / \left(\frac{\partial \mathscr{E}_y}{\partial x}\right) \\ &= \frac{-j\kappa}{\omega \mu_0} A_e \sin(\kappa x) \exp[j(\omega t - \beta z)]\,. \end{aligned} \tag{51}$$

Outside the active layer, the field must decay in order to have guided waves. For $|x| > d/2$, the solutions for the transverse electric field and the longitudinal magnetic field are

$$\mathscr{E}_y(x, z, t) = A_e \cos\left(\frac{\kappa d}{2}\right) \exp\left[-\gamma\left(|x| - \frac{d}{2}\right)\right] \exp[j(\omega t - \beta z)] \tag{52}$$

and

$$\mathscr{H}_z(x, z, t) = \left(\frac{-x}{|x|}\right)\left(\frac{j\gamma}{\omega\mu_0}\right)A_e\cos\left(\frac{\kappa d}{2}\right)\exp\left[-\gamma\left(|x| - \frac{d}{2}\right)\right]\exp[j(\omega t - \beta z)] \tag{53}$$

where

$$\gamma^2 \equiv \beta^2 - \bar{n}_{r1}^2 k_0^2 . \tag{54}$$

Since both κ and γ must be positive real numbers, Eqs. 50 and 54 show that the requirement for guided modes is that $\bar{n}_{r2}k_0^2 > \beta^2$ and $\beta^2 > \bar{n}_{r1}^2 k_0^2$, or

$$\bar{n}_{r2} > \bar{n}_{r1} . \tag{55}$$

This result is identical to Eq. 47.

To determine the separation constant β, we use the boundary condition at the dielectric interface where the tangential component of the magnetic field $\mathscr{H}_z$ must be continuous. From Eqs. 51 and 53, we obtain the eigenvalue equation

$$\tan\left(\frac{\kappa d}{2}\right) = \frac{\gamma}{\kappa} = \sqrt{\frac{\beta^2 - \bar{n}_{r1}^2 k_0^2}{\bar{n}_{r2}^2 k_0^2 - \beta^2}} . \tag{56}$$

The solution for Eq. 56 depends on the argument of the tangent function, which has multiple values with the addition of $2\pi m$ (m is an integer). For $m = 0$, we have the lowest-order or fundamental mode. For $m = 1$, we have the first-order mode, and so on. Once the number is specified, Eq. 56 can be solved numerically or graphically. The result can then be used in Eqs. 49 through 53 for the electric and magnetic fields.

We now define a confinement factor Γ, which is the ratio of the light intensity within the active layer to the sum both within and outside the active layer. Since the light intensity is given by the Poynting vector $\mathscr{E} \times \mathscr{H}$, which is proportional to $|\mathscr{E}_y|^2$, the confinement factor for the symmetrical three-layer dielectric waveguide can be obtained from Eqs. 49 and 52 for the even TE waves:

$$\begin{aligned}\Gamma &= \int_0^{d/2}\cos^2(\kappa x)dx\left\{\int_0^{d/2}\cos^2(\kappa x)dx + \int_{d/2}^{\infty}\cos^2\left(\frac{\kappa d}{2}\right)\exp\left[-2\gamma\left(x - \frac{d}{2}\right)\right]dx\right\}^{-1} \\ &= \left\{1 + \frac{\cos^2(\kappa d/2)}{\gamma[(d/2) + (1/\kappa)\sin(\kappa d/2)\cos(\kappa d/2)]}\right\}^{-1}\end{aligned} \tag{57}$$

Similar expressions may be obtained for odd TE waves as well as for the transverse magnetic (TM) waves. The confinement factor is frequently used because it represents the fraction of the energy of the propagating waves within the active layer.

To date, the most extensively studied heterostructure laser is in the GaAs/Al$_x$Ga$_{1-x}$As system. The energy bandgap of Al$_x$Ga$_{1-x}$As is a function of Al composition. The alloy has direct bandgap up to $x = 0.45$, then becomes an indirect-bandgap semiconductor. For heterostructure lasers, the composition region $0 < x < 0.35$ is of most interest where the direct energy gap can be expressed as[4]

$$E_g(x) = 1.424 + 1.247x \qquad \text{(eV)} . \tag{58}$$

The compositional dependence of the refractive index can be represented by

$$\bar{n}_r(x) = 3.590 - 0.710x + 0.091x^2 . \tag{59}$$

For example, for $x = 0.3$ the bandgap of $Al_{0.3}Ga_{0.7}As$ is 1.798 eV, which is 0.374 eV larger than that of GaAs; and its refractive index of 3.385 is about 6% smaller than that of GaAs.

Figure 20a illustrates the influence of the composition on the optical intensity $|\mathscr{E}_y|^2$ in the direction perpendicular to the junction plane for the three-layer dielectric waveguide $Al_xGa_{1-x}As/GaAs/Al_xGa_{1-x}As$. The curves are calculated from Eqs. 49 and 56 for a wavelength of 0.90 μm (1.38 eV) and for the fundamental mode ($m = 0$). The active-layer thickness d of 0.2 μm is held constant while the composition is varied. A significant increase in confinement occurs when x is increased from 0.1 to 0.2. Figure 20b shows the variation of confinement with d for $x = 0.3$. As the active layer becomes smaller, the light spreads farther into the $Al_{0.3}Ga_{0.7}As$, and less of the total intensity is within the active layer. Confinement is thus less effective. For larger d, where higher-order modes are permitted, Fig. 20c shows that as the mode order increases, more of the light is outside the active region. Therefore, to improve the optical confinement, a lower mode order is preferred.

Figure 21 shows the variation of the confinement factor Γ for the fundamental mode with alloy composition and d. It can be seen that Γ decreases rapidly for $d < \lambda/\bar{n}_{r2}$ ($\approx$ 0.5 μm), where the active-layer thickness becomes less than the wavelength of the radiation. Representing the fraction of the propagating mode within the active layer by Γ is an important concept for understanding the influence of the active-layer thickness on the threshold current density.

12.5 LASER OPERATING CHARACTERISTICS

12.5.1 Device Materials and Structures

Laser Materials. The list of semiconductor materials that exhibit laser action continues to grow. At present, virtually all the lasing semiconductors have direct bandgaps. This is expected since the radiative transition in a direct-bandgap semiconductor is a first-order process (i.e., the momentum is automatically conserved); the transition probability is high. For indirect-bandgap semiconductors, the radiative transition is a second-order process (i.e., it involves phonons or other scattering agents to conserve momentum and energy); thus, the radiative transition is much weaker. Additionally, in indirect-bandgap semiconductors, the free carrier loss due to the injected carrier grows faster with excitation than the gain does.[34]

Figure 22 shows the range of laser emission wavelengths for various semiconductors. The large range covers from near ultraviolet to far infrared. A few materials of choice are worthy of mentioning. GaAs was the first material to lase, and its related $Al_xGa_{1-x}As$ heterojunctions have been most extensively studied, developed, and commercialized. The new class of nitride-based materials ($Al_xGa_{1-x}N$ and $Al_xIn_{1-x}N$) has gone through impressive advancement in the past decade, and has pushed the lower-wavelength limit to $\approx$ 0.2 μm. For the important application of optical-fiber commu-

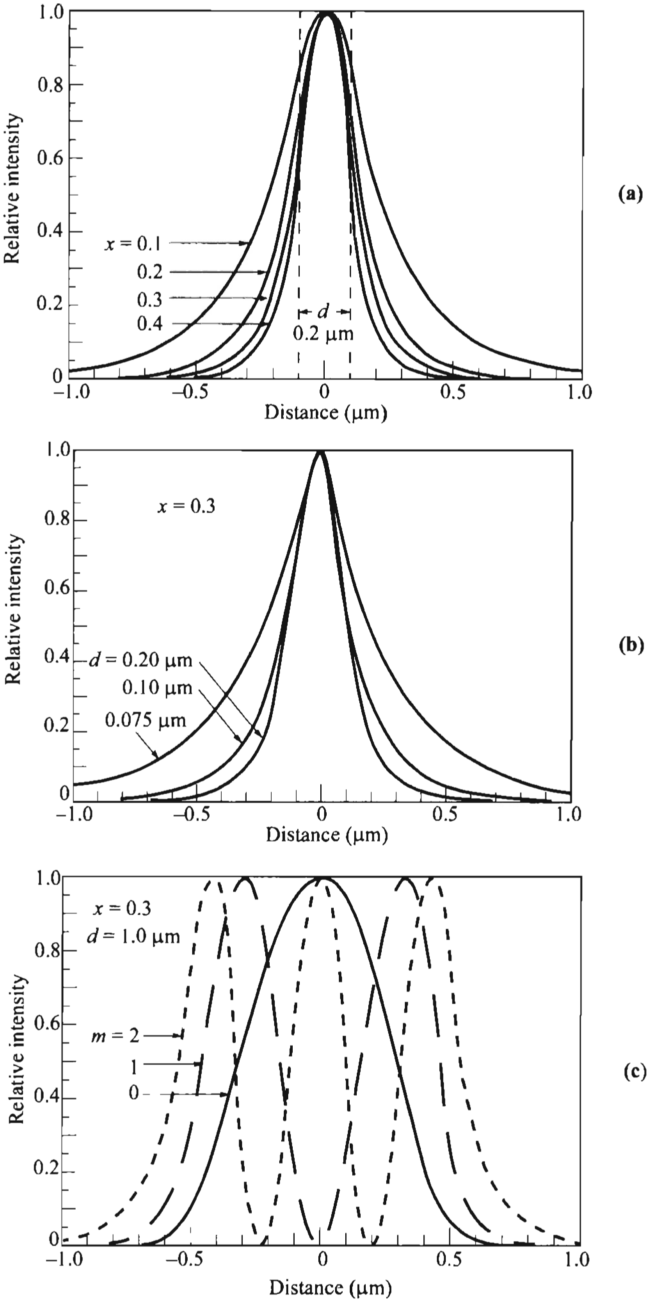


Fig. 20 Square of the electric field (intensity) as a function of position within the double-heterostructure waveguide: (a) For $d = 0.2$ μm and for different AlAs mole fractions. (b) For $x = 0.3$ and different d. (c) For fundamental, first-, and second-order modes with the indicated composition and active-layer thickness. (After Ref. 4.)

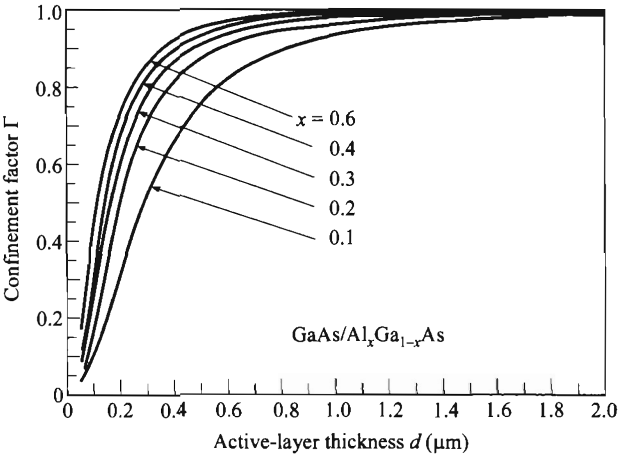


Fig. 21 Confinement factor for fundamental mode as a function of active-layer thickness and alloy composition for a GaAs/$Al_xGa_{1-x}As$ symmetric three-layer dielectric waveguide. (After Ref. 4.)

nication, a wavelength of ≈ 1.55 μm is most desirable (discussed below). This is provided by heterostructures in the $In_xGa_{1-x}As_yP_{1-y}$ and $In_x(Al_yGa_{1-y})_{1-x}As$ systems, and both are lattice matched to the InP substrate. For long-wavelength applications above 3 μm, temperature control below room temperature is required. It should be pointed

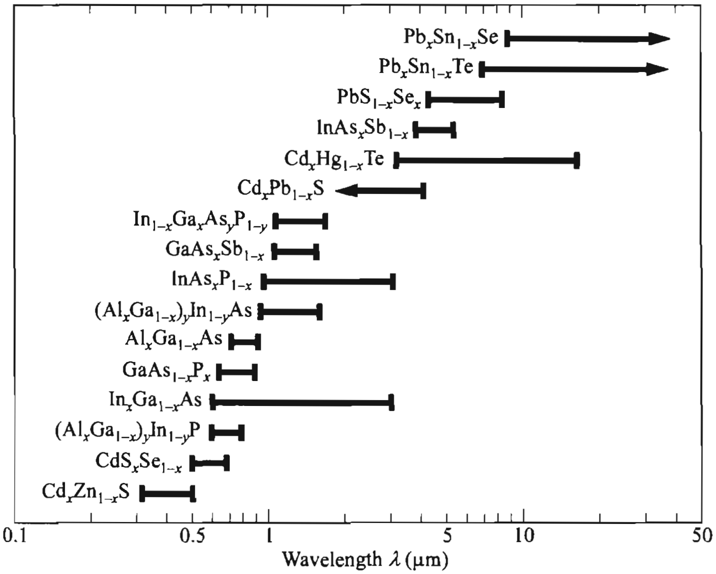


Fig. 22 Laser emission wavelengths for different compound semiconductor materials. (After Ref. 45.)

out that the wavelength ranges indicated are from interband transitions across their bandgaps. For intersubband transitions within the conduction band as in a quantum cascade laser (Section 12.6.3), the wavelength can be much extended for a material system.

Since heterojunction lasers are most common, the bandgap-lattice constant relationship is critical in choosing the appropriate material combination. Such relationships for some common material systems are shown in Fig. 32 of Chapter 1. To achieve heterojunctions with negligible interface traps, the lattices between the two semiconductors must be closely matched. Meanwhile, large bandgap difference is desirable for carrier confinement, and large refractive-index difference is beneficial in waveguiding. Also, as shown in the figure, some of the materials span from direct bandgap to indirect bandgap where lasing is prohibited. Therefore, this composition has to be avoided.

One of the most-important applications for lasers is optical-fiber communication. Figure 23 shows the loss characteristics achieved in experimental fibers. Three wavelengths of particular interest are also indicated on the figure. Around 0.9-μm wavelength, the $GaAs/Al_xGa_{1-x}As$ heterostructure lasers serve as the optical source, and the Si photodiode can be used as the less-expensive photodetector. Around 1.3-μm wavelength the fiber has low loss (0.6 dB/km) and low dispersion; and around 1.55-μm wavelength, the loss reaches a minimum of 0.2 dB/km. For these two wavelengths, III-V quaternary compound lasers, such as $In_xGa_{1-x}As_yP_{1-y}/InP$ lasers, are candidates for optical sources, and photodiodes in ternary or quaternary compounds as well as Ge avalanche photodiodes are candidates for optical detectors.[47]

Device Structures. The basic structure of a laser is a *p-n* junction surrounded by optically designed surfaces, as shown in Fig. 24. A pair of parallel planes are cleaved or polished perpendicular to the plane of the junction. The two remaining sides of the

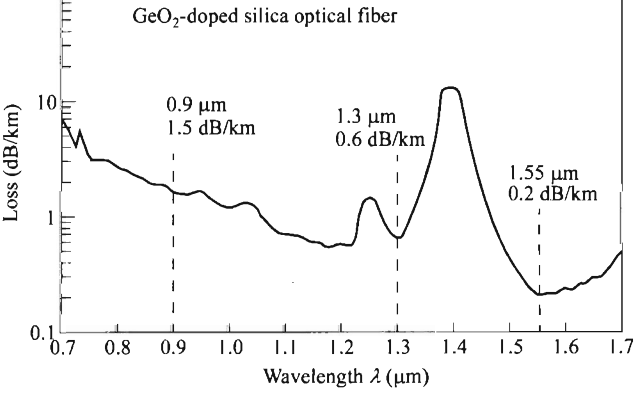


Fig. 23 Loss characteristics of silica optical fiber. The three wavelengths of interest are also shown. (After Ref. 46.)

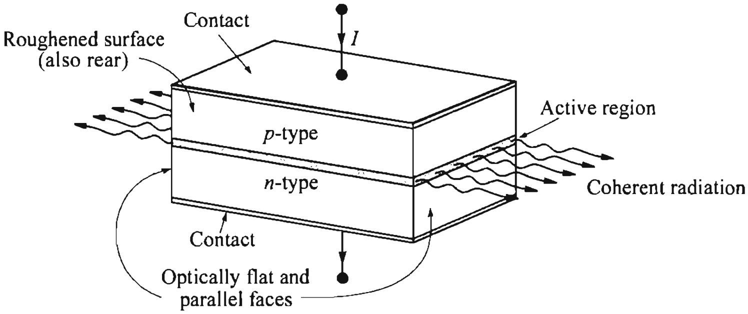


Fig. 24 Basic structure of a junction laser in the form of a Fabry-Perot cavity.

diode are roughened to eliminate lasing in directions other than the main one. The structure is called a Fabry-Perot cavity. When a forward bias is applied to the laser diode, initially at low current there is spontaneous emission. As the bias is increased, eventually a threshold current is reached at which the stimulated emission occurs and a monochromatic and highly directional beam of light is emitted from the junction.

To reduce the threshold current, heterostructure lasers are common practical device structures, using the epitaxial growth techniques. Figure 25 compares a homostructure, a single heterostructure, and a double heterostructure, together with their energy-band diagrams under forward-biased conditions, the refractive-index change, and the optical-field distributions. As can be seen, a single heterostructure can effectively confine the light only at the heterojunction side. But in the double heterostructure (DH), the carriers are confined in the active region d by the heterojunction potential barriers on both sides, and the optical field is also confined within the same active region by the abrupt reduction of the refractive index. These confinements can enhance the stimulated emission and substantially reduce the threshold current. The DH lasers are the most-common configuration.

Several other configurations of heterostructure lasers are of interest.[49] Sometimes it is advantageous to widen the waveguide area while maintaining the area of carrier confinement where light originates. The advantage of such design is greater power output than from regular DH lasers. For example, in a standard DH laser, optical intensity in the waveguide layer can be very high and it sometimes causes catastrophic failure at the reflecting surfaces. Figure 26a shows the separate-confinement heterostructure (SCH) laser which has four heterojunctions. The energy band, the refractive index, and the light intensity perpendicular to the junction plane are plotted. The step in energy between GaAs and $Al_{0.1}Ga_{0.9}As$ is sufficient to confine the carriers within the GaAs layer, but the step in refractive index $\bar{n}_r$ does not sufficiently confine the light. However, the larger step in $\bar{n}_r$ in the outer heterojunctions effectively confines the light and thereby provides the optical waveguide of width W. Low threshold current has been obtained from such a structure.

The large-optical-cavity (LOC) heterostructure laser is similar to a regular DH laser except a p-n homojunction is sandwiched between the two heterojunctions

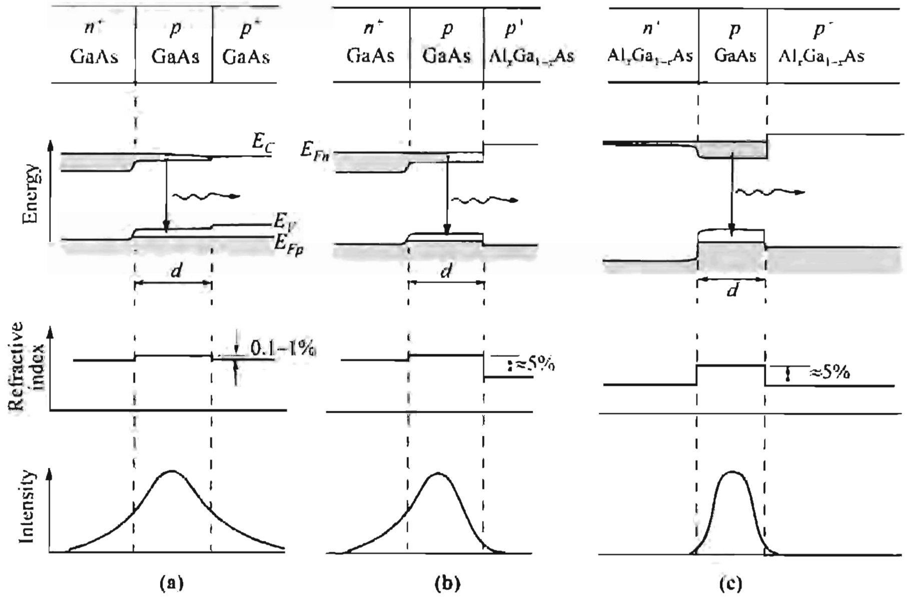
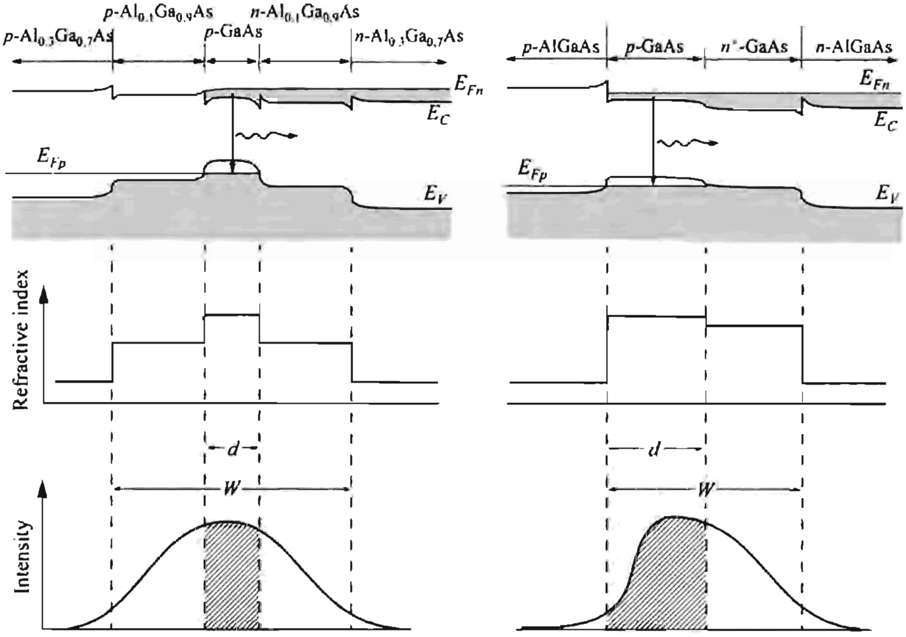


Fig. 25 Comparison of (a) homostructure, (b) single-heterostructure, and (c) double-heterostructure lasers. The top row shows energy-band diagrams under forward bias. n_r change for GaAs/$Al_xGa_{1-x}As$ is about 5% while that due to doping is less than 1%. Confinement of light is shown at bottom. (After Ref. 48.)

(Fig. 26b). Most of the junction current is due to the injection of electrons into the p-layer, which is the active region. The p-GaAs/p-AlGaAs heterojunction provides both carrier and optical confinement, while the n-GaAs/n-AlGaAs heterojunction provides optical confinement only.

The basic laser structures shown so far are broad-area lasers since the whole area along the junction plane can emit radiation. Most heterostructure lasers in practice are made in stripe geometries where the light output is restricted to a narrow beam. The stripe widths are typically 5 to approximately 30 µm. The advantages of the stripe geometry include; (1) achieving fundamental-mode emission along the junction plane (to be discussed later); (2) reduction of the cross-section area which reduces the operation current; (3) improved response time owing to small junction capacitance; and (4) improved reliability by removing most of the junction perimeter from the surface.

Figure 27 shows three representative examples. Methods to restrict current flow to a narrow stripe are called gain-guided. These methods restrict the active region where light emission occurs. Additionally, to confine light propagation after light is produced, a waveguide by refractive-index change helps to maintain a laser beam to a narrow width. This is called index-guided. The three structures in Fig. 27 are all gained-guided. The first is by proton bombardment which produces high-resistivity regions. The lasing area is restricted to the center region which is not bombarded.



(a) (b)

Fig. 26 Schematic representation of energy band (under bias), refractive index, and light intensity for two special heterostructure lasers. (a) Separate-confinement heterostructure (SCH) laser. (b) Large-optical-cavity (LOC) laser. Light emission is within *d* and waveguide is within *W*.

Figure 27b shows a geometry with mesa isolation, formed by etching, and that in Fig. 27c with dielectric isolation. The structure in Fig. 27c also has gone through epitaxial regrowth after mesa etch, with material of higher bandgap and lower refractive index. This structure, thus, is also index-guided by the AlGaAs surrounding material. The structure shows excellent linearity in light-current characteristics and symmetry in laser output from both mirrors.

All the laser structures described above use cavity facets that are formed by cleaving, polishing, or etching to obtain the optical feedback and optical cavity necessary for lasing. Optical feedback can also be provided by a periodic variation of the refractive index within the waveguide, which is generally produced by corrugating the interface between two dielectric layers. The structures in Fig. 28 give two examples. The periodic variation of $\bar{n}_r$ can give rise to constructive interference. Lasers that utilize these corrugated structures are distributed-feedback (DFB) lasers (Fig. 28a) and distributed-Bragg reflector (DBR) lasers (Fig. 28b).[50] The difference between them is the placement of the grating. In the DFB laser, the grating is within the optical cavity of the SCH structure, whereas in the DBR laser, it is outside the active layer. In both of these lasers, the reflection is provided by Bragg reflection in

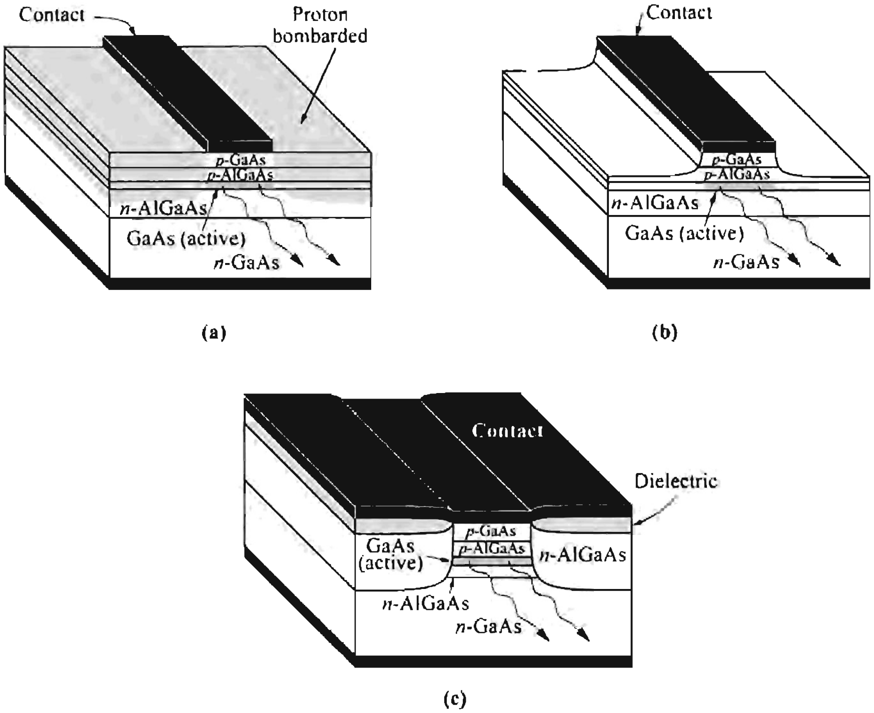

Contact
Proton bombarded
p-GaAs
p-AlGaAs
n-AlGaAs
GaAs (active)
n-GaAs
(a)
Contact
p-GaAs
p-AlGaAs
n-AlGaAs
GaAs (active)
n-GaAs
(b)
Contact
Dielectric
GaAs (active)
p-GaAs
p-AlGaAs
n-AlGaAs
n-AlGaAs
n-GaAs
(c)

Fig. 27 Stripe-geometry DH lasers which are gain-guided with (a) proton isolation, (b) mesa isolation, and (c) dielectric isolation. Structure (c) is also index-guided.

place of the mirrors. The distributed-Bragg reflector is formed by alternating layers of different refractive indexes, with thicknesses equal to a quarter wavelength ($\lambda/4\bar{n}_r$). The DBR has much higher reflectivity than a regular cleaved or etched surface. These heterostructure lasers are useful as sources in integrated optics where cleaving and polishing for mirrors are not possible. Furthermore, the Bragg reflection is a function of wavelength, so tuning is easier to obtain single-mode lasing. Another advantage of these structures is that the operation is less temperature sensitive.[51] The emission wavelength of the Fabry-Perot laser follows the temperature dependence of the energy gap, while that of the DFB and DBR lasers follows the smaller temperature dependence of the refractive index.

12.5.2 Threshold Current

The I-V characteristics of a laser diode are from that of the conventional p-n junction diode (see Chapter 2) and are not discussed here. Even though both sides of the laser junction are highly doped, the level is not as high and the transition region is less abrupt than in a tunnel diode, so there is no negative differential resistance under forward bias.

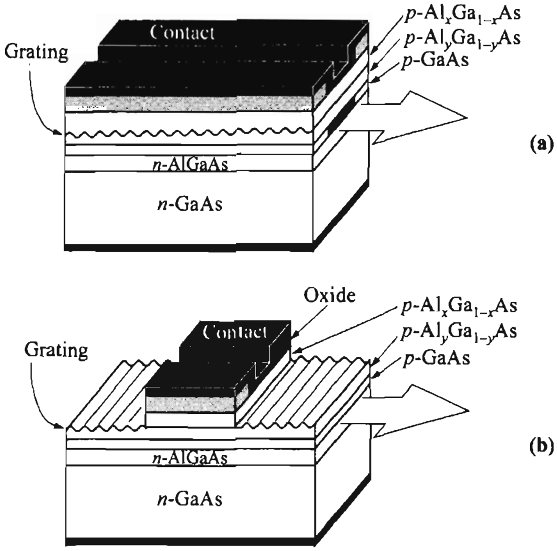

Contact
Grating
p-Al$_x$Ga$_{1-x}$As
p-Al$_y$Ga$_{1-y}$As
p-GaAs
n-AlGaAs
n-GaAs
(a)
Oxide
Contact
Grating
p-Al$_x$Ga$_{1-x}$As
p-Al$_y$Ga$_{1-y}$As
p-GaAs
n-AlGaAs
n-GaAs
(b)

Fig. 28 Schematic structures of (a) distributed-feedback (DFB) laser and (b) distributed-Bragg-reflector (DBR) baser.

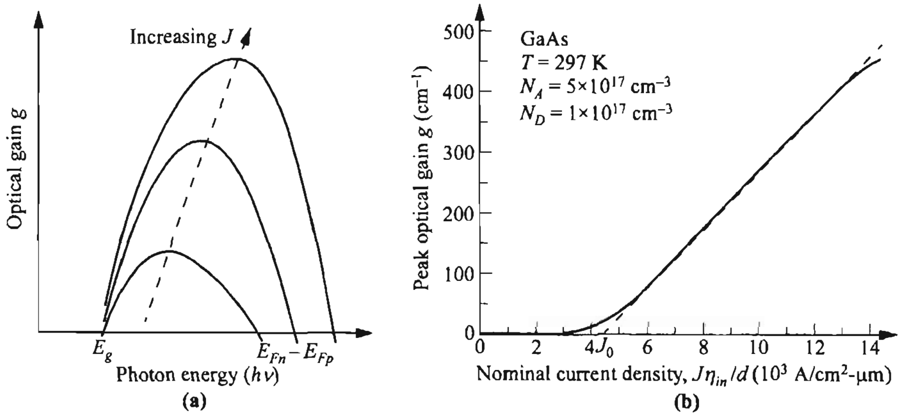


Fig. 29 Optical gain as a function of laser bias current. (a) Optical gain vs. emission photon energy, for different bias current. Range of photon energy reflects Eq. 40. (b) Variation of peak optical gain with nominal current density. (After Ref. 52.)

From the previous discussion, in stimulated emission, the optical gain depends strongly on the electron concentration in the upper energy level. In a laser diode, the injected electron concentration is proportional to the bias current, so the optical gain also has a linear dependence on the bias current. Figure 29 helps to clarify this picture. As the bias current is increased, the Fermi-Dirac distribution functions $F_C(E)$

and $F_V(E)$ change, that is, E_{Fn} increases and E_{Fp} decreases, so the quantity $(E_{Fn} - E_{Fp})$ increases (Fig. 29a). The optical gain increases and the shape of the gain curve also changes. The peak optical gain g shifts to a slightly higher energy (shorter wavelength).

The relationship between optical gain and bias current can be described by the linear equation

$$g = \frac{g_0}{J_0}\left(\frac{J\eta_{in}}{d} - J_0\right). \tag{60}$$

For a nominal current density $(J\eta_{in}/d)$ above a threshold value of J_0, the optical gain increases linearly with the bias current. Figure 29b shows the calculated gain for a sample GaAs laser. The gain is superlinear at low values and increases linearly with J for $50 \le g \le 400$ cm^{-1}. The linear dashed line represents Eq. 60 with $g_0/J_0 = 5\times10^{-2}$ cm-μm/A and $J_0 = 4.5\times10^3$ A/cm^2-μm. At higher current bias, the gain is reduced from the projected value and tends to saturate. This phenomenon of gain saturation arises because for a high rate of stimulated emission, the large extent of population inversion is hard to sustain. A reduced electron concentration in the conduction band results in smaller optical gain until a balance is struck when the supply of carriers can replenish the rate of stimulated emission.

We now address the light output when the bias current is varied. The general characteristics are shown in Fig. 30. At low current there is only spontaneous emission in all directions with a relatively broad spectrum. As the current is increased, the gain increases until the threshold for lasing is reached. The condition for lasing is when the gain is large enough so that a light wave makes a complete traversal of the cavity with the gain equaling the internal loss and external emission. This condition has been discussed previously and given by Eq. 46b. Combining Eqs. 60 and 46b, the threshold current density for lasing is given by[53]

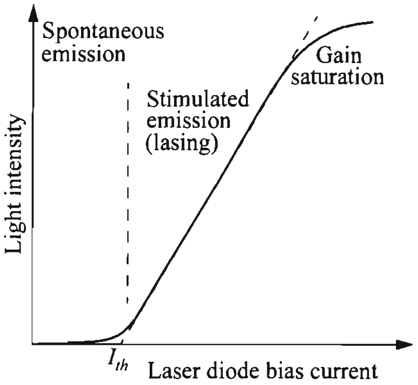


Fig. 30 Light output vs. laser bias current, showing the threshold current.

$$J_{th} = \frac{J_0 d}{\eta_{in}}\left(1 + \frac{g_{th}}{g_0}\right)$$
$$= \frac{J_0 d}{\eta_{in}}\left\{1 + \frac{1}{g_0 \Gamma}\left[\alpha + \frac{1}{2L}\ln\left(\frac{1}{R_1 R_2}\right)\right]\right\}. \tag{61}$$

This equation also takes into account the confinement factor by replacing g_{th} with Γg_{th}. It is seen here in order to reduce the threshold current density, one can increase η_{in}, Γ, L, R_1 and R_2, and to reduce d and α. Achieving low threshold current is one of the main goals in laser development.

Figure 31 compares the calculated J_{th} from Eq. 61 to experimental results.[53] The J_{th} decreases with decreasing d, reaching a minimum, and then increases again. The increase of J_{th} at very narrow active layer thickness is caused by the poor confinement factor Γ. For a given d, J_{th} decreases with increasing Al composition x because of the improved optical confinement. Similar results have been obtained for InP/$Ga_xIn_{1-x}As_yP_{1-y}$/InP DH lasers.[54,55]

The heterostructure lasers have low threshold current density at room temperature because of (1) the carrier confinement provided by the energy barriers of the higher bandgap semiconductor surrounding the active region, and (2) the optical confinement provided by the abrupt reduction of the refractive index outside the active region. Besides threshold currents being lower, they also have less temperature

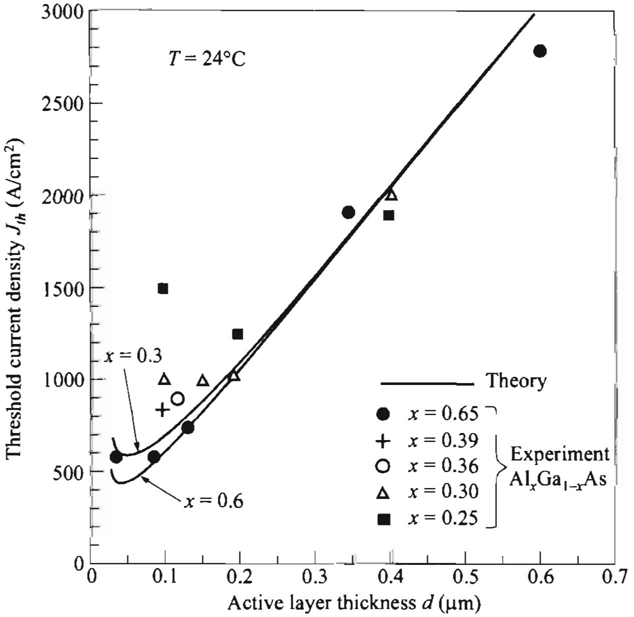


Fig. 31 Comparison of experimental J_{th} and theoretical calculation as a function of d. (After Ref. 53.)

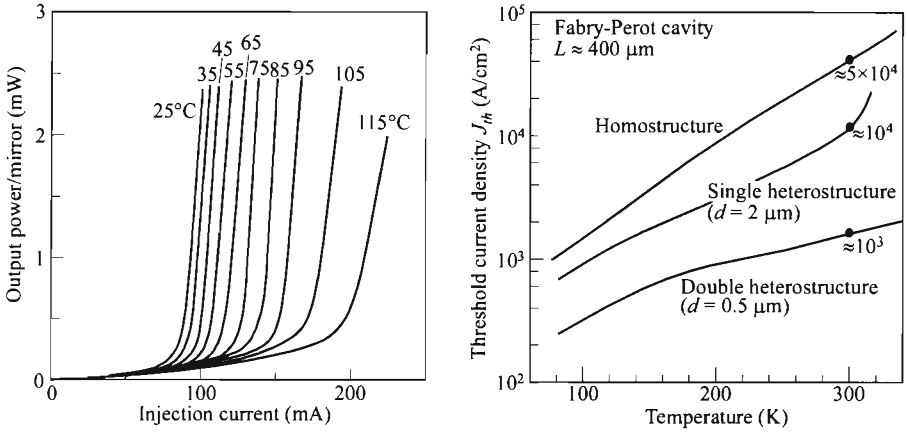
dependence than that in homostructures. Figure 32 shows threshold current versus operating temperature. In DH lasers, The threshold current increases exponentially with temperature as

$$I_{th} \propto \exp\left(\frac{T}{T_0}\right), \tag{62}$$

and T_0 is found to be 110–160°C. Since J_{th} for DH lasers can be less than 10^3 A/cm^2 at 300 K, continuous room-temperature operation is common. This achievement has led to increased applications of semiconductor lasers in science and technology, especially for optical-fiber communication systems. For the homostructure (e.g., GaAs *p-n* junction), the threshold current density J_{th} increases rapidly with increasing temperature. A typical value of J_{th} (obtained by pulse measurement) is about 5.0×10^4 A/cm^2 at room temperature. Such a large current density imposes serious difficulties in operating the laser continuously at 300 K.

12.5.3 Light Spectra and Efficiencies

Figure 33 shows typical output characteristics of semiconductor lasers as the bias current is increased from low currents of spontaneous emission to currents in excess of the laser threshold. At low currents, the spontaneous emission is proportional to the diode bias current, and it has a broad spectral distribution with a typical spectral width at half power of 5 to 20 nm. This is similar to emission in an LED. As the bias current approaches the threshold value, the optical gain can be high enough for amplification so that intensity peaks start to appear. The peaks in wavelength correspond to the



(a) **(b)**

Fig. 32 (a) Light output vs. diode current for GaAs/Al$_x$Ga$_{1-x}$As stripe DH laser at different temperatures, showing temperature dependence of threshold current. (After Ref. 56.) (b) Threshold current density vs. temperature for double-heterostructure, single-heterostructure, and homostructure lasers. (After Ref. 48.)

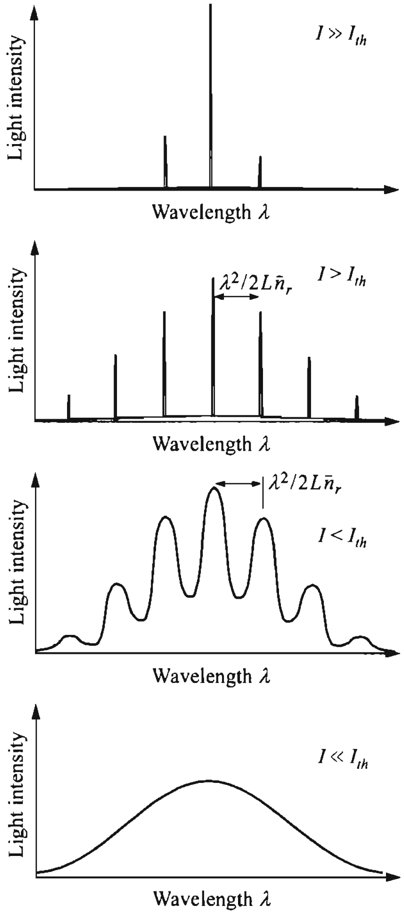


Fig. 33 Emission spectra of laser diode at different bias current: (bottom to top) much below threshold, just below threshold, just above threshold, much above threshold. Scales on intensity increasing from bottom to top.

standing waves in the optical resonator and the spacing between peaks are given by Eq. 43. At this bias level, the light is still incoherent due to the nature of spontaneous emission. When the bias reaches the threshold current, the lasing spectra suddenly become much narrower (< 1 Å), and the light is coherent and much more directional. Also shown are multiple modes lasing simultaneously, called longitudinal modes, but the number of modes can be reduced with further increase of bias current, as shown in the top figure. Also from Eq. 43, the mode spacing is inversely proportional to the cavity length L, so it is advantageous to have small L to limit single-mode operation. This is one of the advantages of semiconductor lasers compared to other laser mediums.

We now consider the power and efficiencies of the laser light output. Above threshold, the power generated by stimulated emission internally is linearly dependent on the bias current,

$$P_{st} = \frac{(I - I_{th})h\nu\eta_{in}}{q}. \tag{63}$$

Referring back to Eq. 46b, the loss per length inside the optical resonator is α, while the average mirror loss from one complete return path is $(1/2L)\ln(1/R_1R_2)$. The power inside the cavity versus output power are proportional to these factors. The laser power output is, thus, given by the ratio of these factors as

$$\begin{aligned} P_{out} &= P_{st}\frac{(1/2L)\ln(1/R_1R_2)}{\alpha + (1/2L)\ln(1/R_1R_2)} \\ &= \frac{(I - I_{th})h\nu\eta_{in}}{q}\left[\frac{\ln(1/R_1R_2)}{2\alpha L + \ln(1/R_1R_2)}\right]. \end{aligned} \tag{64}$$

The external quantum efficiency is defined as the photon emission rate per injected carrier,

$$\begin{aligned} \eta_{ex} &= \frac{d(P_{out}/h\nu)}{d[(I - I_{th})/q]} \\ &= \eta_{in}\left[\frac{\ln(1/R_1R_2)}{2\alpha L + \ln(1/R_1R_2)}\right]. \end{aligned} \tag{65}$$

The overall power efficiency is defined as

$$\eta_P = \frac{P_{out}}{VI} = \frac{(I - I_{th})h\nu\eta_{in}}{VIq}\left[\frac{\ln(1/R_1R_2)}{2\alpha L + \ln(1/R_1R_2)}\right]. \tag{66}$$

In general, the bias qV is slightly higher than the energy gap E_g or photon energy $h\nu$. So η_{in}, η_{ex}, and η_P are very high, in the order of tens of percents.

12.5.4 Far-Field Pattern

The far-field pattern is the intensity profile of the emitted radiation in free space. Due to the small geometry of semiconductor lasers, diffraction causes some degree of divergence of the output beam. Figure 34 gives a schematic representation of the far-field emission of a DH laser. The full angles at half power in the directions perpendicular (θ_x) to and along (θ_y) the junction plane are $\theta_x = \theta_\perp$ and $\theta_y = \theta_{||}$, respectively. For a first-order estimate, the angle is given by the ratio λ/critical dimension. So for a stripe geometry of $d\times S = 1$ µm×10 µm, $\theta_{||}$ is of the order of 10°, whereas $\theta_\perp$ is considerably larger, being 30 to 60°.

The far-field pattern can be calculated by first considering the TE waves in free space for $z > 0$. The wave equation is identical to Eq. 48, except that ε is replaced by ε_0 for free space. Using separation of variables and the boundary condition that $\mathscr{E}_y(x,z)$ must be continuous at $z = 0$, the far-field intensity at an angle θ_x relative to the intensity at $\theta_x = 0$ can be obtained:

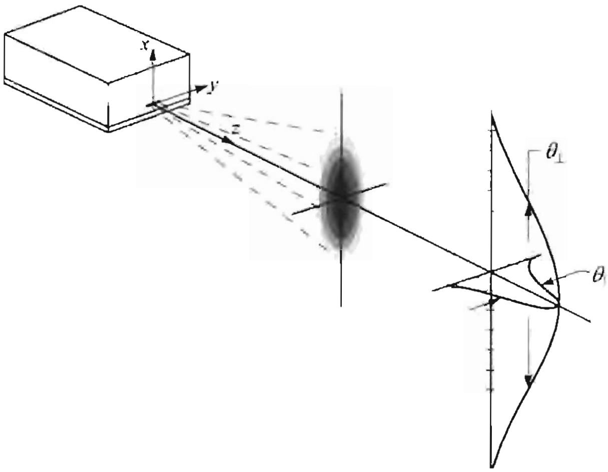


Fig. 34 Schematic representation of far-field emission of a stripe-geometry DH laser. The full angles at half power, perpendicular ($\theta_\perp$) to and along ($\theta_\parallel$) the junction plane, are indicated. (After Ref. 4.)

$$\frac{I(\theta_x)}{I(0)} = \cos^2\theta_x \left| \int_{-\infty}^{\infty} \mathscr{E}_y(x, 0)\exp(j\sin\theta_x k_0 x)dx \right|^2 \times \left| \int_{-\infty}^{\infty} \mathscr{E}_y(x, 0)dx \right|^{-2} . \tag{67}$$

For the symmetrical three-layer waveguide (in DH laser), the electric-field expression of Eqs. 49 and 52 can be substituted into Eq. 67. The full angle $\theta_\perp$ is obtained when the intensity ratio with respect to the maximum is set to 1/2. Figure 35 shows the calculated and measured full angles at half power $\theta_\perp$ for the far-field pattern. The solid

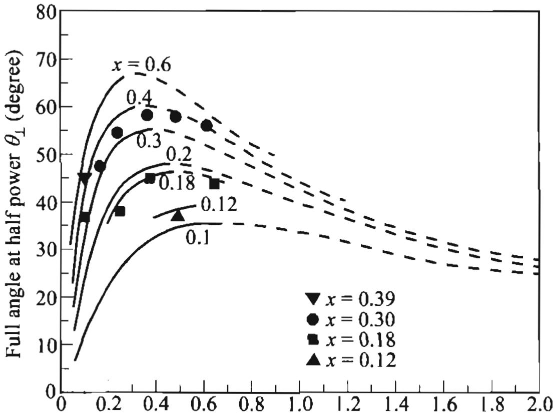


Fig. 35 Theoretical and experimental (symbols) full angle at half power $\theta_\perp$ as a function of active-layer thickness and composition of GaAs/Al$_x$Ga$_{1-x}$As DH lasers. (After Ref. 57.)

curves are the beam divergence calculated from Eq. 67 for the fundamental mode. The dashed portion represents a range of active-layer thickness where high-order modes are possible. The experimental data points are in good agreement with the calculations. For a typical active-layer thickness of 0.2 μm in a GaAs/$Al_{0.3}Ga_{0.7}As$ DH laser, the full angle $\theta_\perp$ is about 50°.

The electric-field intensity of the stripe-geometry lasers along the direction parallel to the junction plane (y-direction) is strongly influenced by the spatial variation of the dielectric permittivity. For the stripe structure shown in Fig. 36, the wave equation with a sinusoidal time dependence given by $\exp(j\omega t)$ is[58]

$$\nabla^2 \mathscr{E}_y + \frac{k_0^2 \varepsilon}{\varepsilon_0} \mathscr{E}_y = 0 . \tag{68}$$

In this equation, k_0 equals $2\pi/\lambda$ and $\varepsilon/\varepsilon_0$ is taken as two-dimensional, with the form

$$\frac{\varepsilon(x,y)}{\varepsilon_0} = \frac{\varepsilon(0) - a^2 y^2}{\varepsilon_0} \tag{69}$$

to simulate an index-guided active layer, and

$$\frac{\varepsilon(x,y)}{\varepsilon_0} = \frac{\varepsilon_1}{\varepsilon_0} \tag{70}$$

in the adjacent inactive layers. In Eq. 69, $\varepsilon(0)$ is the complex dielectric permittivity $\varepsilon_r(0) + j\varepsilon_i(0)$ at $y = 0$ in the active layer, and a is a complex constant represented by $a_r + ja_i$. An approximate solution of Eq. 68 having dielectric permittivity given by Eqs. 69 and 70 is

$$\mathscr{E}_x(x,y,z) = \mathscr{E}_y(x)\mathscr{E}_y(y)\exp(-j\beta_z z) . \tag{71}$$

Since $\varepsilon(x,y)$ varies slowly with y along the junction plane, $\mathscr{E}_y(x)$ is not significantly affected by the confinement along y and can be represented by the previously derived expressions, Eqs. 49 and 52. From Eq. 68, by separation of variables one obtains

$$\frac{\partial^2 \mathscr{E}_y(x)}{\partial x^2} + \beta_x^2 \mathscr{E}_y(x) = 0 . \tag{72}$$

Substituting Eqs. 71 and 72 into Eq. 68 and eliminating $\mathscr{E}_y(x)$ by multiplying its complex conjugate, and integrating over x yields a differential equation for $\mathscr{E}_y(y)$:

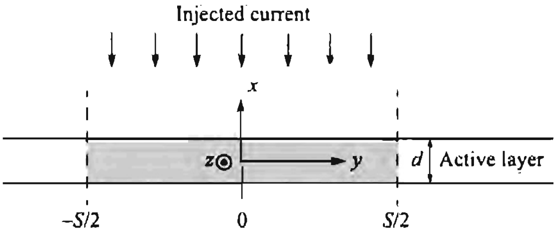


Fig. 36 Coordinate system for a stripe-geometry laser with active layer d and stripe width S.

$$\frac{\partial^2 \mathscr{E}_y(y)}{\partial y^2} + \left\{ k_0^2 \left[\frac{\Gamma \varepsilon(0)}{\varepsilon_0} + (1-\Gamma)\frac{\varepsilon_1}{\varepsilon_0} \right] - \beta_x^2 - \beta_z^2 - \frac{\Gamma k_0^2 a^2 y^2}{\varepsilon_0} \right\} \mathscr{E}_y(y) = 0 . \qquad (73)$$

The field distributions for $\mathscr{E}_y(y)$ represented by Eq. 73 are Hermite-Gaussian functions given by

$$\mathscr{E}_y(y) = H_p\left(y \sqrt{\frac{\Gamma^{1/2} a k_0}{\varepsilon_0^{1/2}}} \right) \exp\left(-\frac{1}{2} \sqrt{\frac{\Gamma}{\varepsilon_0}} a k_0 y^2 \right) , \qquad (74)$$

where H_p is the Hermite polynomial of order p, which is given by

$$H_p(\xi) \equiv (-1)^p \exp(\xi^2) \frac{\partial^p \exp(-\xi^2)}{\partial \xi^p} . \qquad (75)$$

The first three Hermite polynomials are $H_0(\xi) = 1$, $H_1(\xi) = 2\xi$, and $H_2(\xi) = 4\xi^2 - 2$. Therefore, the intensity for the fundamental mode is Gaussian and is given by

$$|\mathscr{E}_y(y)|^2 = \exp\left[-\sqrt{\frac{\Gamma}{\varepsilon_0}} a_r k_0 y^2 \right] \qquad (76)$$

which demonstrates that the intensity distribution along the junction plane is influenced by a_r.

Figure 37 shows the far-field patterns along the junction plane for stripe-geometry lasers. For a stripe width of 10 μm, a fundamental Gaussian mode distribution exists. As the stripe width increases, higher-order modes along the junction plane are observed. These modes are characteristic of the Hermite-Gaussian distribution represented by Eq. 74. The results show that even though $\theta_{||}$ is reduced for larger stripe width, multiple lobes appear. So the overall beam size and divergence is smaller for smaller stripe width.

12.5.5 Turn-On Delay and Modulation Response

One of the many advantages of a semiconductor laser is that it can be turned on and off with the bias current. This is especially important for high-speed applications such as that in optical-fiber communication. When a current step above the threshold value is applied to a laser, a delay of a few nanoseconds generally occurs before the stimulated emission is observed. The delay time t_d is related to the minority-carrier lifetimes. Also if the bias current is modulated with a small ac signal, the light intensity follows the waveform but only to a certain frequency limit. Both of these put a limit on the frequency response.

To derive the delay time, we consider the continuity equation for electrons in a p-type semiconductor. Under the conditions that the current density J is uniform across the active layer d, and the injected electron concentration n is much greater than the thermal equilibrium value, the continuity equation becomes

$$\frac{dn}{dt} = \frac{J}{qd} - \frac{n}{\tau} - \frac{c g N_{ph}}{\bar{n}_r} , \qquad (77)$$

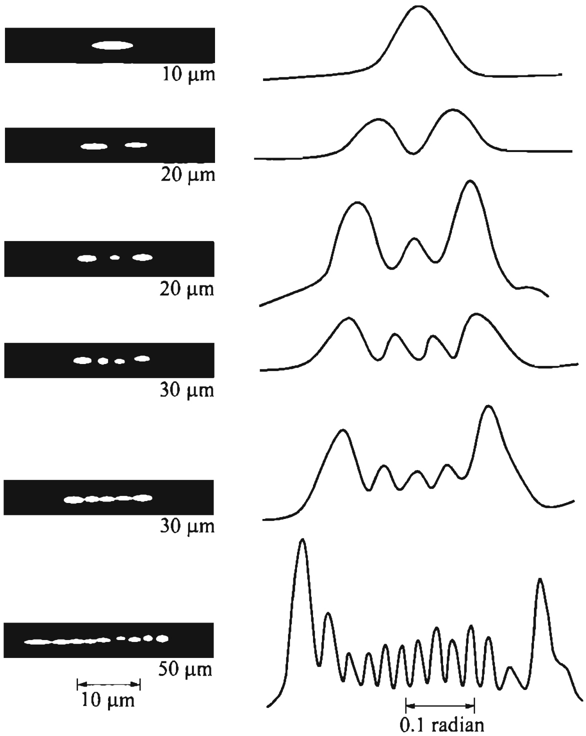


Fig. 37 Near-field (left) and far-field (right) patterns along the junction plane (y-direction) for different stripe width S of DH lasers. (After Ref. 59.)

where τ is the carrier lifetime (Eq. 24), and N_{ph} is the photon density. The first term on the right is the uniform injection rate, the second is the spontaneous recombination rate, and the last the stimulated-emission recombination rate. A similar expression may be written for holes in the n-side of the active layer. For consideration of the turn-on delay time, the last term can be ignored. The solution of this equation, with the initial condition of $n(0) = 0$, is

$$n(t) = \frac{\tau J}{qd}\left[1 - \exp\left(\frac{-t}{\tau}\right)\right] \tag{78a}$$

or

$$t = \tau \ln\left[\frac{J}{J - qn(t)d/\tau}\right]. \tag{78b}$$

When $n(t)$ reaches the threshold value for stimulated emission, the electron concentration also has a threshold value of $n(t) = n_{th}$, related to the threshold current by

$$J_{th} = \frac{qn_{th}d}{\tau}. \tag{79}$$

Since $t = t_d$ at $n(t) = n_{th}$, the turn-on delay time is then

$$t_d = \tau \ln\left(\frac{J}{J - J_{th}}\right). \tag{80}$$

If the laser is prebiased to a current level $J_0 < J_{th}$, solving Eq. 77 with initial condition $n(0) = J_0 \tau / qd$ gives a reduced delay time of

$$t_d = \tau \ln\left(\frac{J - J_0}{J - J_{th}}\right). \tag{81}$$

Figure 38 shows the measured results on the variation of the laser turn-on delay with current, for active layers with different acceptor concentrations. The delay time t_d varies logarithmically in accordance with Eq. 80; delay time decreases as N_A becomes larger, from shorter carrier lifetime.

We next consider the frequency response of the laser output when the bias current is modulated with an ac frequency. The continuity equation for photons is given by

$$\frac{dN_{ph}}{dt} = \frac{cgN_{ph}}{\bar{n}_r} - \frac{N_{ph}}{\tau_{ph}} \tag{82}$$

where N_{ph} is the internal photon density which is proportional to the output light intensity. The spontaneous emission is ignored in this equation. The photon lifetime τ_{ph} is given by

$$\tau_{ph} = \frac{\bar{n}_r}{c[\alpha + (1/2L)\ln(1/R_1R_2)]} \tag{83}$$

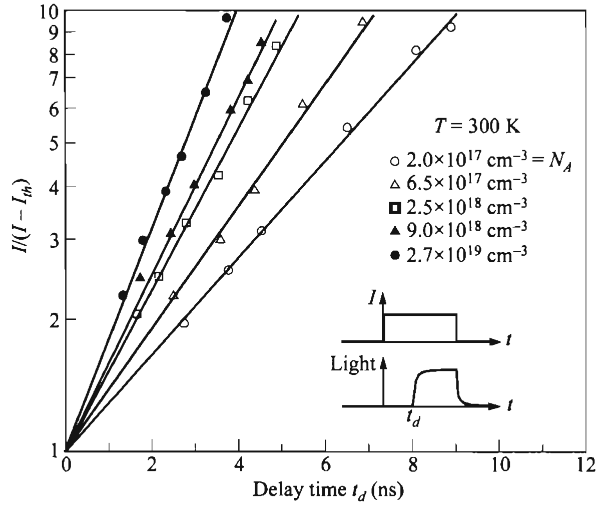


Fig. 38 Variation of laser turn-on delay with current. The delay time t_d is indicated in the insert. (After Ref. 60.)

and is the average lifetime in the cavity before the photons are lost from absorption or emission through the two mirrors. The solution of Eq. 82 has the form[61]

$$\frac{\Delta N_{ph}}{\Delta J} = \frac{\tau}{qd}\left[\left(1 - \frac{f^2}{f_r^2}\right)^2 + (2\pi f \tau_{ph})^2\right]^{-1/2}, \tag{84}$$

where ΔN_{ph} and ΔJ are the small-signal values, and the resonance frequency or so-called relaxation oscillation frequency is given by

$$f_r = \frac{1}{2\pi}\sqrt{\frac{1}{\tau\tau_{ph}}\left(\frac{J_0}{J_{th}} - 1\right)}. \tag{85}$$

The simple form of Eq. 84 indicates the frequency response of laser light. At low frequencies, the response is flat. It has a peak at f_r, above which the response drops quickly according to f^{-2}, or 40 dB per decade of frequency. The f_r or overall response can also be pushed to a higher frequency range with higher dc bias current.

For optical-fiber communications, the optical source must be able to be modulated at high frequencies. DH lasers have good modulation characteristics well into the GHz range. Figure 39 shows a normalized modulated light output as a function of the modulation frequency for an InGaAsP/InP DH laser diode. The laser diode emitting at 1.3 μm is directly modulated with sinusoidal current superposed on the dc bias current. The overall shape and trend of Eqs. 84 and 85 are observed.

Another effect, called frequency chirp, sets in at high frequencies. This originates from the fact that the refractive index in the active layer is changed by the injected carrier concentration, so the refractive index is also modulated to some extent. This causes a shift in emission frequency from that of the dc value.

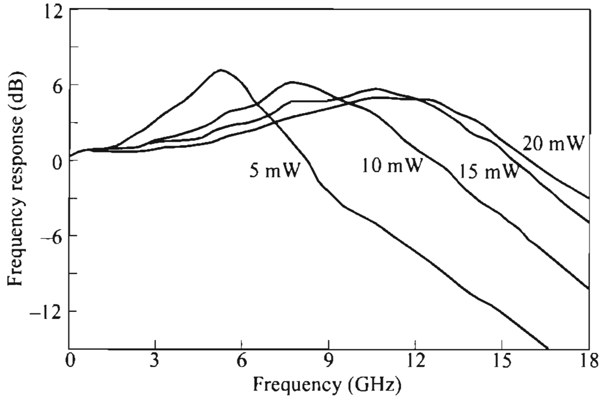


Fig. 39 Normalized small-signal response vs. modulation frequency, at different powers for an InGaAsP distributed-feedback laser, at room temperature. (After Ref. 62.)

12.5.6 Wavelength Tuning

The wavelength ranges covered by compound semiconductor lasers are shown in Figure 22. By choosing the appropriate material and composition for one of the compounds, a laser can be produced at any desired wavelength in the broad range 0.2 to over 30 μm. The emission wavelength of a semiconductor laser can be also varied by varying the diode current or heat-sink temperature, or by applying a magnetic field or pressure.[63]

Bias current level changes the emission wavelength because the injected carrier concentration changes the refractive index of the cavity, and it also shifts the peak photon energy according to Eq. 40 and as indicated in Fig. 17c. The main temperature dependence comes from the change in bandgap energy. Figure 40 shows temperature tuning of a DH PbTe/$Pb_{1-x}Sn_xTe$ laser. By changing the heat-sink temperature from 10 to 120 K, the emission wavelength can be varied from 16 μm to 9 μm.

Applying hydrostatic pressure to a laser diode can provide a broad tuning range. Pressure affects emission wavelength through its effects on (1) the energy gap, (2) the cavity length, and (3) the refractive index. The bandgap varies linearly with hydrostatic pressure for some binary compounds (e.g., InSb, PbS, and PbSe). A PbSe laser at 77 K can be tuned from 7.5 μm to 22 μm using hydrostatic pressure up to 14 kbars.[63]

Diode lasers can also be tuned by a magnetic field. For semiconductors with large effective mass anisotropy, the magnetic energy levels depend on the orientation of the applied magnetic field with respect to the crystal axis. Both conduction and valence bands have their energies quantized into Landau levels. As the magnetic field

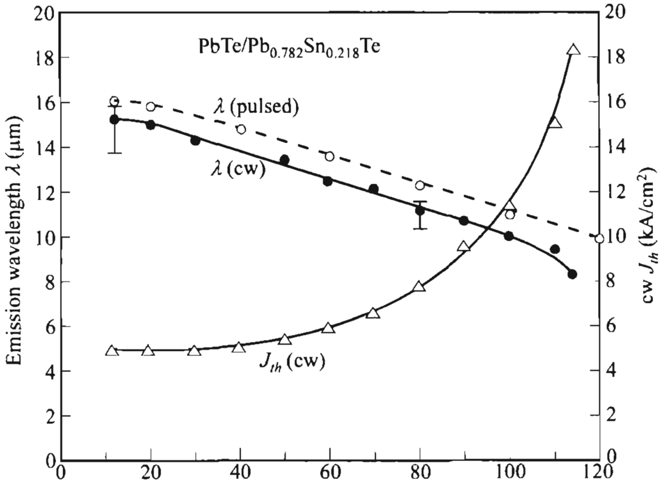


Fig. 40 Variation of emission wavelength and threshold current density as a function of temperature. (After Ref. 64.)

increases, the energy separations between available transitions also increase, causing the emission wavelength to decrease. For example, in a $Pb_{0.79}Sn_{0.21}Te$ laser at 7 K in a ⟨100⟩ magnetic field, the wavelength decreases from 15 μm to 14 μm when a magnetic field of 10 kG is applied.

12.5.7 Laser Degradation

Injection lasers can degrade by a variety of mechanisms. The three main mechanisms are catastrophic degradation, dark-line defect formation, and gradual degradation.[4]

For catastrophic degradation, the laser mirror under high-power operation is permanently damaged by pits or grooves forming on the mirror. The problem is worse with initial flaws on the facets, and it can be improved with some special coating such as Al_2O_3. Modifications of the device structures that reduce surface recombination and absorption increase the power possible at the damage limit.[65]

The dark-line defect is a network of dislocations that can form during laser operation and intrudes upon the optical cavity. Once started, it can grow expansively in a few hours. It is a form of nonradiative recombination centers, causing the threshold current to increase. The progression of these defects bas been linked to the initial material quality. To reduce the probability of dark-line defect formation, quality epitaxial layers grown on substrates with low dislocation density should be used, and the laser should be carefully bonded to the heat sink to minimize strain.

By eliminating instantaneous catastrophic failure and the fairly rapid degradation caused by dark-line defect formation, DH lasers have a long operating life with relatively slow degradation. DH GaAs/AlGaAs lasers under cw operation longer than 3 years at 30°C have not shown signs of degradation.[66] The extrapolated lifetime at 22°C heat-sink temperature is more than 100 years. It is reasonable to assume that the lifetime obtained for GaAs DH lasers can also be achieved operating at longer wavelengths. Similar observations had been made on GaInAsP/InP DH lasers.[47] The long life will meet the requirements for large-scale optical-fiber communication systems as well as satisfying the requirements for other applications.

12.6 SPECIALTY LASERS

12.6.1 Quantum-Well, Quantum-Wire, and Quantum-Dot Lasers

Quantum-Well Laser. When the active-layer thickness of a double-heterostructure (DH) laser is reduced to the order of the de Broglie wavelength ($\lambda = h/p$), two-dimensional quantization occurs and results in a series of discrete energy levels given by the bound-state energies of a finite square well. Such devices are called quantum-well lasers.[67,68] Some basic characteristics of the quantum well have been discussed in Section 1.7, and readers are referred to that section. The advantages of quantum-well lasers are low threshold current, high quantum efficiency, high output power, low temperature dependence, high speed, and wider range of wavelength tuning.

Figure 41a shows the quantum-well potential for a $GaAs/Al_xGa_{1-x}As$ heterostructure, where the well thickness L_x is of the order of 10 nm. The energy eigenvalues are

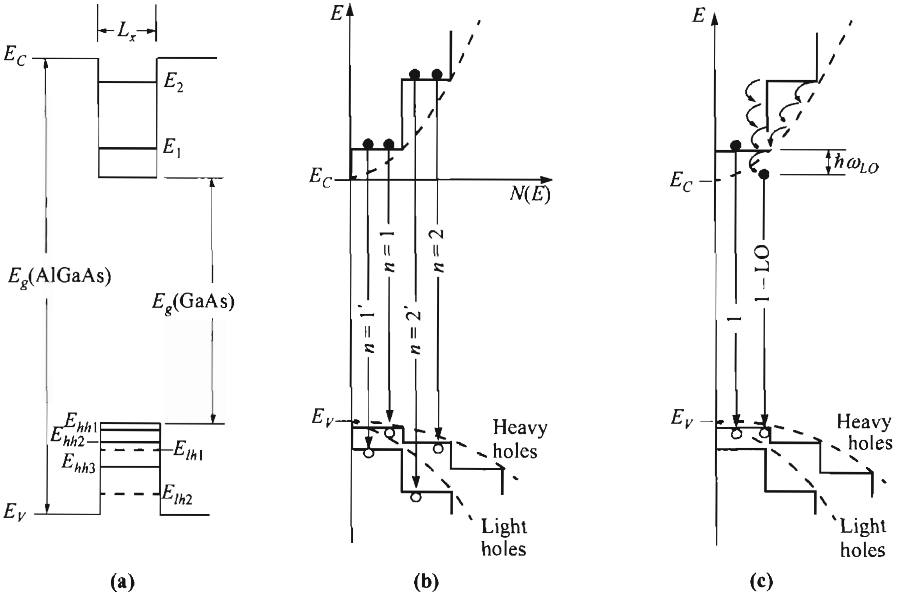


Fig. 41 (a) Potential of quantum well and quantized levels. (b) Density-of-states diagram and possible recombinations. (c) Phonon-assisted recombination in quantum well. (After Ref. 67.)

designated by E_1, E_2 for electrons; by E_{hh1}, E_{hh2}, E_{hh3} for heavy holes; and by E_{lh1}, E_{lh2} for light holes. These quantized levels from their respective band edges are inversely proportional to L_x^2. Figure 41b shows the corresponding density-of-state diagram. The half-parabolas (dashed lines) that originate from the band edges correspond to the densities of states of the bulk semiconductor. The step-like densities of states are characteristics of a quantum-well structure. Interband recombination transitions ($\Delta n = 0$ selection rule) occur from a bound state in the conduction band (say at E_1) to a bound state in the valence band (say at E_{hh1}). The energy of the transition is given by

$$h\nu = E_g(\text{GaAs}) + E_1 + E_{hh1} . \tag{86}$$

The recombination can thus proceed between two well-defined energy levels that are varied by tailoring the quantum-well thickness.

Figure 41c shows another important feature of quantum-well heterostructures. Carriers injected at higher energy can generate phonons and scatter downward in energy, ultimately to a lesser density of states. In a bulk semiconductor, phonon generation is limited by the decreasing density of state, particularly at the band edge; whereas in a quantum-well system, within a constant density-of-states region, there is no such limitation. The photon energy is reduced by the amount of the longitudinal optical-phonon energy $\hbar\omega_{LO}$. This process can transfer an electron to well below the confined-particle states, for example, below E_1 (Fig. 41c). If this amount is larger than E_1 itself, it can lead to laser operation at energies $h\nu < E_g$, instead of, as expected without phonon participation, $h\nu > E_g$ in the usual case.

Many of the advantages of the quantum-well laser arise from the unique shape of the density of states in a 2-D system. The reduction of threshold current is explained as follows, apart from the fact that the active-layer thickness is thin (Eq. 61). Figure 42 compares the density of states in 3-D (bulk) and 2-D (quantum-well) systems, as well as their electron concentration distribution. In a 3-D system, since the density of states varies as $\sqrt{E}$ and approaches zero at the band edge, the electron distribution, obtained by multiplying the Fermi-Dirac distribution function, has a wide spread in energy level. In a quantum well (2-D), the density of states is constant within each subband. Consequently the electron profile at the band edge, E_1 in this case, is much sharper. This condition makes population inversion much easier to achieve, thus lowering the threshold current.

At high current bias, more than one subbands are filled with injected carriers. The emission spectrum internally is thus much wider. The lasing wavelength, however, is also selected by other means such as the optical cavity length. So in a quantum-well laser, wavelength tuning can cover a wider range. This is in additional to the variation of quantum-well width which controls the quantization levels.

One drawback of the thin active layer in a quantum-well laser is poor optical confinement. This can be improved with multiple quantum wells stacked on top of one another. Multiple-quantum-well lasers have higher quantum efficiency as well as higher output power. Single or multiple quantum wells can be incorporated in a separate-confinement heterostructure (SCH) scheme to improve optical confinement.

When the separations of the multiple quantum wells are reduced to the same order as the well thickness, a *superlattice laser* is formed. Minibands start to appear in both the conduction band and the valence band in the active superlattice region. Stimulated emission is from transitions between these minibands.

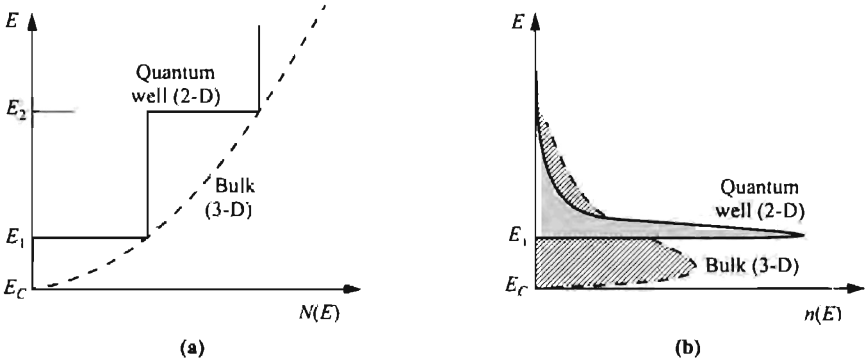


Fig. 42 Comparison of 3-D and 2-D systems: (a) Density of states in conduction band, and (b) electron concentration distribution.

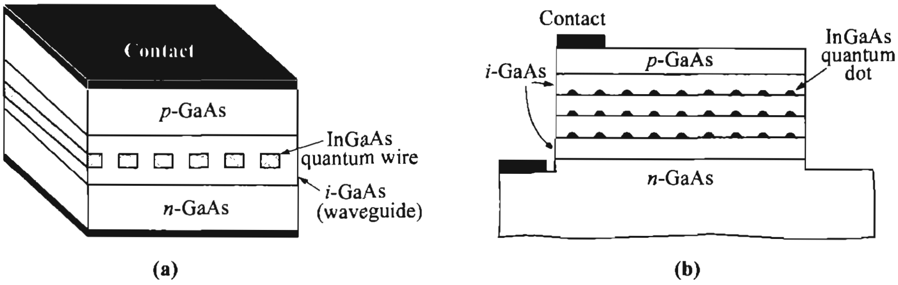


Fig. 43 Simplified schematic structures for (a) quantum-wire laser and (b) quantum-dot laser.

Quantum-Wire and Quantum-Dot Lasers. In quantum-wire and quantum-dot lasers, the active regions are reduced to the de Broglie wavelength regime, into 1-D (wire) and 0-D (island) formation.[69] These wires and dots are placed between a *p-n* junction as shown in Fig. 43. To realize such small dimensions, the small active regions are mostly formed by epitaxial regrowth on specially processed surfaces (etched, cleaved, vicinal, or V-groove), or by a process called self-ordering after epitaxy.[70] The advantages of these lasers are similar to the quantum-well laser, except to a higher degree. These advantages also stem from their respective densities of states, and readers are referred to Section 1.7. These densities of states give rise to the optical-gain spectrums which are compared in Fig. 44. These optical gains include those from a regular 3-D (bulk) active layer down to quantum dots. As seen, the peak gains for quantum wires and quantum dots are progressively higher, and their shapes also sharper. These gain characteristics give the aforementioned desirable advantages

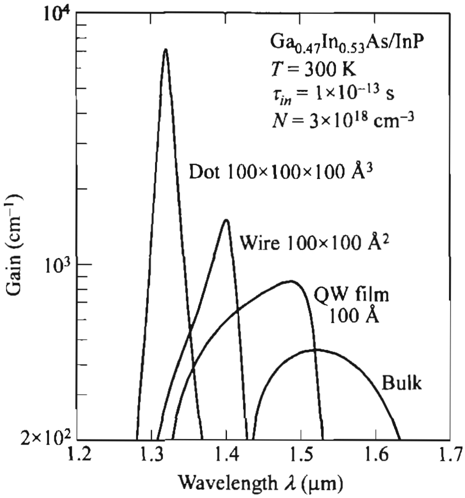


Fig. 44 Calculated optical gain vs. wavelength for different dimensionality. Note increase of peak gain and narrower spectrum as dimensionality is reduced. (After Ref. 71.)

such as low threshold current. The reduction of threshold current for different structures are summarized in Fig. 45, which also indicates their introduction in chronological order.

12.6.2 Vertical-Cavity Surface-Emitting Laser (VCSEL)

So far the lasers discussed are edge-emitting such that the light output is parallel to the active layer. In a surface-emitting laser, the light output is orthogonal to the active layer (heterointerfaces) and the semiconductor surface. Note that the optical cavity is now defined by planes parallel to the heterointerfaces, as shown in Fig. 46, so the name *vertical-cavity surface-emitting laser* (VCSEL).[73,74] The optical cavity is formed by two distributed-Bragg reflectors (DBRs) surrounding the active layer. These DBRs have high reflectivity larger than 90%. The high reflectivity is required since the optical gain per pass is small due to the small optical cavity compared to an edge-emitting laser. The VCSEL usually has an active layer formed by multiple quantum wells. The benefits of a small optical cavity include low threshold current, and single-mode lasing since the mode separation is wide (Eq. 43). Other advantages of the VCSEL are the realization of 2-D laser array, ease of coupling light output to other mediums such as optical fiber and optical interconnect, compatibility with IC processing for integrated optics, high-volume and low-cost production, high speed, on-wafer testing capability, etc.

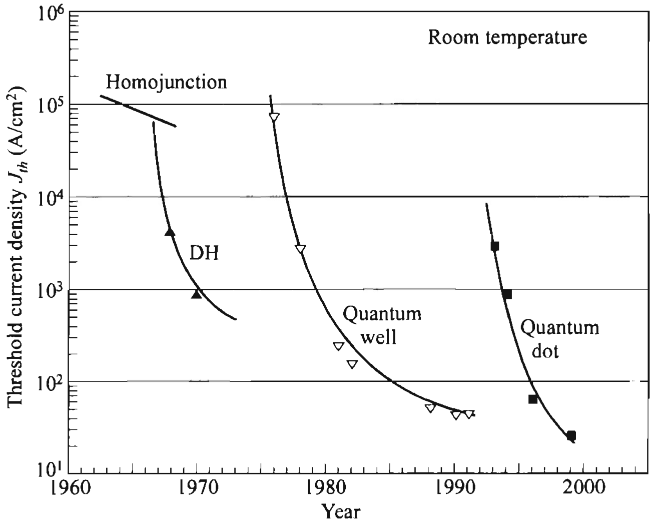


Fig. 45 Reduction of threshold current density from homojunction laser to DH, quantum-well, and quantum-dot lasers. (After Ref. 72.)

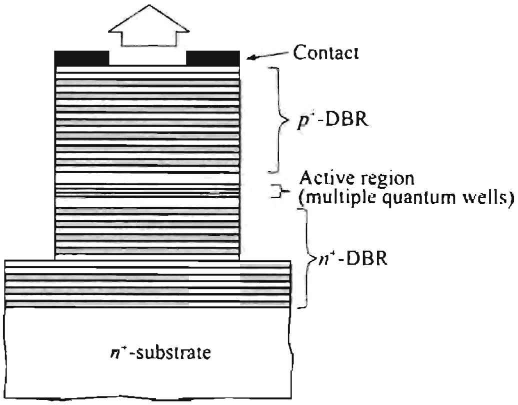


Fig. 46 Structure of vertical-cavity surface-emitting laser (VCSEL).

12.6.3 Quantum-Cascade Laser

In a *quantum cascade laser*, the electron transition to emit a photon is between quantized subband energy levels, created by a quantum well or superlattice, within the same conduction band (Fig. 47).[75] The major difference is *intersubband* transition as opposed to interband transition in a regular laser. Since the transition between subbands is much smaller than the energy gap, the quantum cascade laser is capable of lasing in long wavelengths, without facing the material difficulties of very narrow energy gap, which are much less stable and less developed. Wavelengths beyond 70 μm have been achieved. Besides, the wavelength is tunable by the quantum-well thickness without being fixed by the energy gap.

The active region is composed of multiple quantum wells or a superlattice. The most-common design is between two to three quantum wells. In the active region, electrons are injected through resonant tunneling, to the sublevel E_3. (Readers are referred to Section 8.4 for resonant tunneling.) The radiative transition between E_3 and E_2 is responsible for lasing. Electrons in E_2 relax to E_1 and then tunnel to the miniband of the succeeding injector through resonant tunneling, or they can also

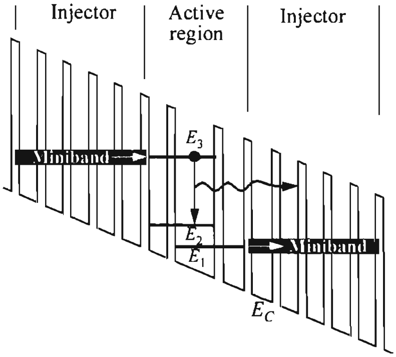


Fig. 47 Energy-band diagram showing conduction-band edge E_C of the quantum cascade laser under lasing condition. A period consists of an active region and a superlattice injector, and is repeated in series.

tunnel directly from E_2 to the injector. Resonant tunneling is a very fast process so that the concentration in E_2 is always less than that in E_3; thus population inversion is maintained. Design of the minibands plays a critical role and it depends on the nonuniform thicknesses of the quantum wells. Notice that E_3 is not aligned to a miniband of the succeeding injector, so tunneling to the injector is blocked and high concentration at E_3 can be sustained.

The design of the injector is also critical. Under bias, the miniband should remain flat for efficient resonant tunneling. This has to be done by careful tailoring of the injector superlattice with a special doping profile, thickness profile, or barrier profile.

The period, consisting of the active region plus injector, is repeated many times (20–100), and this *cascade* scheme helps to establish a high external quantum efficiency and low threshold current since the same carrier can produce many photons. This phenomenon is not possible in a conventional laser. Also due to the small transition energy, lower-temperature operation is inevitable. Nevertheless, CW operation has been achieved at ≈ 150 K, and pulsed operation has been made possible at room temperature.

12.6.4 Semiconductor Optical Amplifier

The *semiconductor optical amplifier* (SOA), sometimes called a *semiconductor laser amplifier*, is very similar to a laser except that its mirrors of the optical resonator are much less reflecting, so there are very few internal optical passes.[76] It can be thought as a laser but operated below its threshold current so that an additional input light (as optical pumping) is needed to start the stimulated emission process, resulting in an optical gain over the input light signal. There are two types of SOA: *Fabry-Perot* or *resonant SOA* and *traveling-wave SOA*. The difference again lies in the mirror reflectivity. A Fabry-Perot SOA has a medium mirror reflectivity of approximately 30%. Its spectrum has longitudinal modes similar to that of Fig. 33 in a regular laser. The traveling-wave SOA has very low mirror reflectivity ($<10^{-4}$) and it assumes a single optical pass, so it does not have multiple modes as in a Fabry-Perot SOA. The advantage of the Fabry-Perot SOA is higher gain, but it may suffer from reaching gain saturation and sometimes lasing, both of which can be circumvented in a traveling-wave SOA.

The SOA is useful in an optical-fiber communication system as an in-line optical amplifier or repeater. It is a much simpler device that can replace a system that requires a photodetector, an electrical amplifier, and a laser.

REFERENCES

1. H. J. Round, "A Note on Carborundum," *Electrical World*, **49**, 309 (1907).
2. A. A. Bergh and P. J. Dean, *Light-Emitting Diodes*, Clarendon, Oxford, 1976.
3. H. F. Ivey, "Electroluminescence and Semiconductor Lasers," *IEEE J. Quantum Electron.*, **QE-2**, 713 (1966).
4. H. C. Casey, Jr. and M. B. Panish, *Heterostructure Lasers*, Academic, New York, 1978.

5. M. G. Craford, "Recent Developments in LED Technology," *IEEE Trans. Electron Dev.*, **ED-24**, 935 (1977).
6. E. F. Schubert, *Light-Emitting Diodes*, Cambridge University Press, Cambridge, 2003.
7. W. N. Carr, "Characteristics of a GaAs Spontaneous Infrared Source with 40 Percent Efficiency," *IEEE Trans. Electron Dev.*, **ED-12**, 531 (1965).
8. P. Goldberg, Ed., *Luminescence of Inorganic Solids*, Academic, New York, 1966.
9. C. H. Gooch, *Injection Electroluminescent Devices*, Wiley, New York, 1973.
10. S. Wang, *Solid-State Electronics*, McGraw-Hill, New York, 1966.
11. P. C. Eastman, R. R. Haering, and P. A. Barnes, "Injection Electroluminescence in Metal-Semiconductor Tunnel Diodes," *Solid-State Electron.*, **7**, 879 (1964).
12. O. V. Lossev, *Wireless World Radio Rev.*, **271**, 93 (1924).
13. O. V. Lossev, "Luminous Carborundum Detector and Detection Effect and Oscillations with Crystals," *Philos. Mag.*, **6**, 1024 (1928).
14. J. R. Haynes and H. B. Briggs, "Radiation Produced in Germanium and Silicon by Electron-Hole Recombination," *Bull. Am. Phys. Soc.*, **27**, 14 (1952).
15. R. J. Keyes and T. M. Quist, "Recombination Radiation Emitted by Gallium Arsenide," *Proc. IRE*, **50**, 1822 (1962).
16. J. I. Pankove and J. E. Berkeyheiser, "A light Source Modulated at Microwave Frequencies," *Proc. IRE*, **50**, 1976 (1962).
17. J. I. Pankove and M. J. Massoulie, "Injection Luminescence from Gallium Arsenide," *Bull. Am. Phys. Soc.*, **7**, 88 (1962).
18. H. G. Grimmeiss and H. Scholz, "Efficiency of Recombination Radiation in GaP," *Phys. Lett.*, **8**, 233 (1964).
19. A. C. Eten and J. H. Haanstra, "Electroluminescence in Tellurium-Doped Cadmium Sulphide," *Phys. Lett.*, **11**, 97 (1964).
20. D. G. Thomas, J. J. Hopfield and C. J. Frosch, "Isoelectronic Traps due to Nitrogen in Gallium Phosphide," *Phys. Rev. Lett.*, **15**, 857 (1965).
21. S. Nakamura, "III-V Nitride-Based LEDs and Lasers: Current Status and Future Opportunities," *Tech. Dig. IEEE IEDM*, 9 (2000).
22. W. O. Groves, A. H. Herzog, and M. G. Craford, "The Effect of Nitrogen Doping on GaAsP Electroluminescent Diodes," *Appl. Phys. Lett.*, **19**, 184 (1971).
23. L. S. Rohwer and A.M. Srivastava, "Development of Phosphors for LEDs," *Interface*, 36, (summer 2003).
24. J. E. Geusic, F. W. Ostermayer, H. M. Marcos, L. G. Van Uitert, and J. P. Van Der Ziel, "Efficiency of Red, Green and Blue Infrared-to-Visible Conversion Sources," *J. Appl. Phys.*, **42**, 1958 (1971).
25. U. Kaufmann, M. Kunzer, K. Köhler, H. Obloh, W. Pletschen, P. Schlotter, J. Wagner, A. Ellens, W. Rossner, and M. Kobusch, "Single Chip White LEDs," *Phys. Stat. Sol.*, (a), **192**, 246 (2002).
26. K. Ikeda, S. Horiuchi, T. Tanaka, and W. Susaki, "Design Parameters of Frequency Response of GaAs-AlGaAs DH LED's for Optical Communications," *IEEE Trans. Electron Dev.*, **ED-24**, 1001 (1977).
27. J. P. Gordon, H. J. Zeiger, and C. H. Townes, "Molecular Microwave Oscillator and New Hyperfine Structure in the Microwave Spectrum of NH_3," *Phys. Rev.*, **95**, 282 (1954).

28. N. G. Basov and A. M. Prokhorov, "Application of Molecular Beams to the Radio Spectroscopic Study of the Rotation Spectra of Molecules," *Zh. Eksp. Theo. Fiz.*, **27**, 431 (1954).
29. T. H. Maiman, "Stimulated Optical Radiation in Ruby Masers," *Nature* (*Lond.*), **187**, 493 (1960).
30. P. Aigrain (1958), as reported in *Proc. Conf. Quantum Electron.*, Paris, 1963, p. 1762.
31. N. G. Basov, B. M. Vul, and Y. M. Popov, "Quantum-Mechanical Semiconductor Generators and Amplifiers of Electromagnetic Oscillations," *Sov. Phys. JEPT*, **10**, 416 (1960).
32. W. S. Boyle and D. G. Thomas, U.S. Patent 3,059,117 (Oct. 16, 1962, filed Jan. 1960).
33. M. G. A. Bernard and G. Duraffourg, "Laser Conditions in Semiconductors," *Phys. Status Solidi*, **1**, 699 (1961).
34. W. P. Dumke, "Interband Transitions and Maser Action," *Phys. Rev.*, **127**, 1559 (1962).
35. R. N. Hall, G. E. Fenner, J. D. Kingsley, T. J. Soltys, and R. O. Carlson, "Coherent Light Emission from GaAs Junctions," *Phys. Rev. Lett.*, **9**, 366 (1962).
36. M. I. Nathan, W. P. Dumke, G. Burns, F. H. Dill, Jr., and G. J. Lasher, "Stimulated Emission of Radiation from GaAs *p-n* Junction," *Appl. Phys. Lett.*, **1**, 62 (1962).
37. T. M. Quist, R. H. Rediker, R. J. Keyes, W. E. Krag, B. Lax, A. L. McWhorter, and H. J. Zeigler, "Semiconductor Maser of GaAs," *Appl. Phys. Lett.*, **1**, 91 (1962).
38. N. Holonyak, Jr. and S. F. Bevacqua, "Coherent (Visible) Light Emission from $Ga(As_{1-x}P_x)$ Junctions," *Appl. Phys. Lett.*, **1**, 82 (1962).
39. H. Kroemer, "A Proposed Class of Heterojunction Injection Lasers," *Proc. IEEE*, **51**, 1782 (1963).
40. Z. I. Alferov and R. F. Kazarinov, U.S.S.R. Patent 181,737. Filed 1963. Granted1965.
41. I. Hayashi, M. B. Panish, P. W. Foy, and S. Sumski, "Junction Lasers which Operate Continuously at Room Temperature," *Appl. Phys. Lett.*, **17**, 109 (1970).
42. R. N. Hall, "Injection Lasers," *IEEE Trans. Electron Dev.*, **ED-23**, 700 (1976).
43. A. L. Schawlow, "Masers and Lasers," *IEEE Trans. Electron Dev.*, **ED-23**, 773 (1976).
44. I. Hayashi, "Heterostructure Lasers," *IEEE Trans. Electron Dev.*, **ED-31**, 1630 (1984).
45. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, New York, 1991.
46. T. Miya, Y. Terunuma, T. Hosaka, and T. Miyashita, "Ultimate Low-Loss Single Mode Fiber at 1.55 μm," *Electron. Lett.*, **15**, 108 (1979).
47. A. G. Foyt, "1.0–1.6 μm Sources and Detectors for Fiber Optics Applications," *IEEE Device Res. Conf.*, Boulder, Colo., June 25, 1979.
48. M. B. Panish, I. Hayashi, and S. Sumski, "Double-Heterostructure Injection Lasers with Room Temperature Threshold as Low as 2300 A/cm^2," *Appl. Phys. Lett.*, **16**, 326 (1970).
49. C. A. Burrus, H. C. Casey, Jr., and T. Y. Li, "Optical Sources," in S. E. Miller and A. G. Chynoweth, Eds., *Optical Fiber Communication*, Academic, New York, 1979.
50. H. C. Casey, Jr., S. Somekh, and M. Ilegems, "Room-Temperature Operation of Low-Threshold Separate-Confinement Heterostructure Injection Laser with Distributed Feedback," *Appl. Phys. Lett.*, **27**, 142 (1975).
51. K. Aiki, M. Nakamura, and J. Umeda, "Lasing Characteristics of Distributed-Feedback GaAs-GaAlAs Diode Lasers with Separate Optical and Carrier Confinement," *IEEE J. Quantum Electron.*, **QE-12**, 597 (1976).
52. F. Stern, "Calculated Spectral Dependence of Gain in Excited GaAs," *J. Appl. Phys.*, **47**, 5382 (1976).

53. H. C. Casey, Jr., "Room Temperature Threshold-Current Dependence of GaAs-$Al_xGa_{1-x}As$ Double Heterostructure Lasers on x and Active-Layer Thickness," *J. Appl. Phys.*, **49**, 3684 (1978).

54. R. E. Nahory and M. A. Pollack, "Threshold Dependence on Active-Layer Thickness in InGaAsP/InP DH Lasers," *Electron. Lett.*, **14**, 727 (1978).

55. M. Yana, H. Nishi, and M. Takusagawa, "Theoretical and Experimental Study of Threshold Characteristics in InGaAsP/InP DH Lasers," *IEEE J. Quantum Electron.*, **QE-15**, 571 (1979).

56. W. T. Tsang, R. A. Logan, and J. P. Van der Ziel, "Low-Current-Threshold Stripe-Buried-Heterostructure Lasers with Self-Aligned Current Injection Stripes," *Appl. Phys. Lett.*, **34**, 644 (1979).

57. H. C. Casey, Jr., M. B. Panish, and J. L. Merz, "Beam Divergence of the Emission from Double-Heterostructure Injection Lasers,"*Appl. Phys. Lett.*, **44**, 5470 (1973).

58. T. L. Paoli, "Waveguiding in a Stripe-Geometry Junction Laser," *IEEE J. Quantum Electron.*, **QE-13**, 662 (1977).

59. H. Yonezu, I. Sakuma, K. Kobayashi, T. Kamejima, M. Ueno, and Y. Nannichi, "A GaAs-$Al_xGa_{1-x}As$ Double Heterostructure Planar Stripe Laser," *Jpn. J. Appl. Phys.*, **12**, 1585 (1973).

60. C. J. Hwang and J. C. Dyment, "Dependence of Threshold and Electron Lifetime on Acceptor Concentration in GaAs-$Ga_{1-x}Al_xAs$ Lasers," *J. Appl. Phys.*, **44**, 3240 (1973).

61. P. Bhattacharya, *Semiconductor Optoelectronic Devices*, 2nd Ed., Prentice Hall, Upper Saddle River, New Jersey, 1997.

62. N. K. Dutta, S. J. Wang, A. B. Piccirilli, R. F. Karlicek, Jr., R. L. Brown, M. Washington, U. K. Chakrabarti, and A. Gnauck, "Wide-Bandwidth and High-Power InGaAsP Distributed Feedback Lasers," *J. Appl. Phys.*, **66**, 4640 (1989).

63. I. Melngailis and A. Mooradian, "Tunable Semiconductor Diode Lasers and Applications," in S. Jacobs, M. Sargent, J. F. Scott, and M. O. Scully, Eds., *Laser Applications to Optics and Spectroscopy*, Addison-Wesley, Reading, Mass., 1975.

64. J. N. Walpole, A. R. Calawa, T. C. Harman, and S. H. Groves, "Double-Heterostructure PbSnTe Lasers Grown by Molecular-Beam Epitaxy with CW Operation up to 114 K," *Appl. Phys. Lett.*, **28**, 552 (1976).

65. H. Yonezu, I. Sakuma, T. Kamojima, M. Ueno, K. Iwamoto, I. Hino, and I. Hayashi, "High Optical Power Density Emission from a Window Stripe AlGaAs DH Laser," *Appl. Phys. Lett.*, **34**, 637 (1979).

66. R. L. Hartman, N. E. Schumaker, and R. W. Dixon, "Continuously Operated AlGaAs DH Lasers with 70°C Lifetimes as Long as Two Years," *Appl. Phys. Lett.*, **31**,756 (1977).

67. N. Holonyak, Jr., R. M. Kolbas, R. D. Dupuis, and P. D. Dapkus, "Quantum-Well Heterostructure Lasers," *IEEE J. Quantum Electron.*, **QE-16**, 170 (1980).

68. B. Zhao and A. Yariv, "Quantum Well Semiconductor Lasers," in *Semiconductor Lasers I: Fundamentals*, E. Kapon, Ed., Academic Press, San Diego, CA, 1999.

69. E. Kapon, "Quantum Wire and Quantum Dot Lasers," in *Semiconductor Lasers I: Fundamentals*, E. Kapon, Ed., Academic Press, San Diego, CA, 1999.

70. J. M. Moison, F. Houzay, F. Barthe, L. Leprince, E. André, and O. Vatel, "Self-Organized Growth of Regular Nanometer-Scale InAs Dots on GaAs," *Appl. Phys. Lett.*, **64**, 196 (1994).

71. M. Asada, Y. Miyamoto, and Y. Suematsu, "Gain and the Threshold of Three-Dimensional Quantum-Box Lasers," *IEEE J. Quantum Electron.*, **QE-22**, 1915 (1986).
72. N. N. Ledentsov, M. Grundmann, F. Heinrichsdorff, D. Bimberg, V. M. Ustinov, A. E. Zhukov, M. V. Maximov, Z. I. Alferov, and J. A. Lott, "Quantum-Dot Heterostructure Lasers," *IEEE J. Selected Topics Quan. Elect.*, **6**, 439 (2000).
73. K. D. Choquette, "Vertical-Cavity Surface-Emitting Lasers: Light for the Information Age," *MRS Bulletin*, 507, (July 2002).
74. J. M. Rorison, "Vertical Cavity Surface Emitting Lasers for Communications," in B. Krauskopf and D. Lenstra, Eds., *Fundamental Issues of Nonlinear Laser Dynamics*, American Inst. Phys., 2000.
75. F. Capasso, R. Paiella, R. Martini, R. Colombelli, C. Gmachl, T. L. Myers, M. S. Taubman, R. M. Williams, C. G. Bethea, K. Unterrainer, H. Y. Hwang, D. L. Sivco, A. Y. Cho, A. M. Sergent, H. C. Liu, and E. A. Whittaker, "Quantum Cascade Lasers: Ultrahigh-Speed Operation, Optical Wireless Communication, Narrow Linewidth, and Far-Infrared Emission," *IEEE J. Quantum Electron.*, **QE-38**, 511 (2002).
76. N. A. Olsson, "Semiconductor Optical Amplifiers," *Proc. IEEE*, **80**, 375 (1992).

PROBLEMS

1. The spectrum for spontaneous emission is given by Eq. 6. Find (a) the photon energy at the peak of the spectrum and (b) the spectrum width (i.e. the half-power width).
2. Find the spectrum width in wavelength for the spontaneous emission. If the central wave length is in the middle of the visible spectrum (0.555 μm), what is the spectrum width at room temperature?
3. Assume that the radiative lifetime τ_r is given by $\tau_r = 10^9/N$ seconds, where N is the semiconductor doping in cm^{-3} and the nonradiative lifetime τ_{nr} is equal to 10^{-7} s. Find the cutoff frequency of an LED having a doping of 10^{19} cm^{-3}.
4. A GaAs sample is illuminated with a light having a wavelength of 0.6 μm. The incident power is 15 mW. If one-third of the incident power is reflected and another third exits from the other end of the sample, what is the thickness of the sample? Find the thermal energy dissipated per second to the lattice.
5. An InGaAsP Fabry-Perot laser operating at a wavelength of 1.3 μm has a cavity length of 300 μm. The refractive index of InGaAsP is 3.39.
   a) What is the mirror loss expressed in cm^{-1}?
   b) If one of the laser facets is coated to produce 90% reflectivity, how much threshold current reduction (as a percentage) can be expected, assuming $\alpha = 10$ cm^{-1}?
6. a) For an InGaAsP laser operating at a wavelength of 1.3 μm, calculate the mode spacing in nanometer for a cavity of 300 μm, assuming the group refractive index is 3.4.
   b) Express the mode spacing obtained above in GHz.
7. The confinement factor can be approximated by $\Gamma = 1 - \exp(-C\Delta\bar{n}d)$ where C is a constant, $\Delta\bar{n}$ is the difference in the refractive indexes, and d is the thickness of the active layer. If $C = 8\times10^5$ cm^{-1}, $d = 1$ μm, the refractive index of GaAs to be 3.6, and a critical angle at the active-to-nonactive boundary of 78° (between GaAs and AlGaAs double heterojunction), find the confinement factor.

8. Calculate the threshold current for the case in Prob. 7 if one end-mirror's reflectivity is 0.99. The cavity width is 5 μm, the lose per unit length $\alpha = 100$ cm^{-1}, and the gain factor is 0.1 cm^{-3}A^{-1} [gain factor $\equiv (J_0 d/\eta_{in} g_0 L)^{-1}$].

9. If the refractive index is dependent on the wavelength, find the separation $\Delta\lambda$ between the allowed modes in the longitudinal direction. For a GaAs laser diode operated at $\lambda = 0.89$ μm, with $\bar{n}_r = 3.58$, $L = 300$ μm, and $d\bar{n}_r/d\lambda = 2.5$ μm^{-1}, what is $\Delta\lambda$?

10. The temperature dependence of the threshold current can be expressed as $I_{th} = I_0 \exp(T/T_0)$, and the temperature coefficient is $\xi \equiv (1/I_{th})(dI_{th}/dT)$. For high-temperature operation, it is important to have a low ξ. What is the coefficient ξ for the laser shown in Fig. 32a? If $T_0 = 50$°C, is this laser better or worse for high-temperature operation?

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

13

Photodetectors and Solar Cells

13.1 INTRODUCTION
13.2 PHOTOCONDUCTOR
13.3 PHOTODIODES
13.4 AVALANCHE PHOTODIODE
13.5 PHOTOTRANSISTOR
13.6 CHARGE-COUPLED DEVICE (CCD)
13.7 METAL-SEMICONDUCTOR-METAL PHOTODETECTOR
13.8 QUANTUM-WELL INFRARED PHOTODETECTOR
13.9 SOLAR CELL

13.1 INTRODUCTION

Photodetectors are semiconductor devices that can detect optical signals through electronic processes. The extension of wavelength of coherent and incoherent light sources into the far-infrared region on one hand and the ultraviolet region on the other has increased the need for high-speed, sensitive photodetectors. The operation of a general photodetector includes basically three processes: (1) carrier generation by incident light, (2) carrier transport and/or multiplication by current-gain mechanism if present, and (3) extraction of carriers as terminal current to provide the output signal.

Photodetectors are important in optical-fiber communication systems operated in the near-infrared region (0.8 to 1.6 μm). They demodulate optical signals, that is, convert the optical variations into electrical variations, that are subsequently amplified and further processed. For such applications the photodetectors must satisfy stringent requirements such as high sensitivity at operating wavelengths, high response speed, and minimum noise. In addition, the photodetector should be compact in size, use low biasing voltage and current, and be reliable under operating conditions.

There exist many types of photodetectors. They are divided into two classes; thermal detectors and photon detectors. Thermal detectors detect light by sensing the temperature rise when the light energy is absorbed at their dark surface. These are more suitable for far-infrared wavelengths. Technically they are more like thermal sensors and will be discussed to a larger extent in the next chapter. The photon detectors are based on the quantum photoelectric effect: a photon excites a carrier which contributes to the photocurrent. This chapter only discusses semiconductor photon detectors which are majority of the photodetectors in the commercial market.

The large variety of photodetectors, as indicated by the chapter sections, are required because they target different aspects of performance. To understand the advantages of each photodetector, we discuss the performance metric of a photodetector. Since the photoelectric effect is based on the photon energy $h\nu$, the wavelength of interest is related to energy transition ΔE in the device operation, with the obvious but important relationship

$$\lambda = \frac{hc}{\Delta E} = \frac{1.24}{\Delta E(\text{eV})} \quad (\mu\text{m}) \tag{1}$$

where λ is the wavelength, c is the speed of light, and ΔE is the transition of energy levels. Since usually photon energy $h\nu > \Delta E$ can also cause excitation, Eq. 1 is often the minimum wavelength limit for detection. The transition energy ΔE, in most cases, is the energy gap of the semiconductor. But depending on the type of photodetector, it can be the barrier height as in a metal-semiconductor photodiode, or transition energy between an impurity level and the band edge as in an extrinsic photoconductor. The type of photodetector and the semiconductor material are chosen and optimized for the wavelength of interest.

The absorption of light in a semiconductor is indicated by the absorption coefficient. Not only does it determine whether light can be absorbed for photoexcitation, but it also indicates where light is absorbed. A high value of absorption coefficient indicates light is absorbed near the surface where light enters. A low value means the absorption is low that light can penetrate deeper into the semiconductor. In the extreme, light can be transparent for long wavelengths without photoexcitation. It thus determines the quantum efficiency of a photodetector. Figure 1 shows the measured intrinsic absorption coefficients for various photodetector materials.[1] The solid curves are for 300 K and the dashed curves for 77 K. For Ge, Si, and III-V compound semiconductors, the curves shift toward longer wavelengths as the temperature increases. For some IV-VI compounds (e.g., PbSe), the opposite happens as the bandgap increases with increasing temperature. (The emission wavelengths of some important lasers are also shown for reference.)

The speed of photodetectors is important, especially for optical-fiber communication systems. The response of the photodetector has to be fast enough compared to the digital transmission data rate, where light is turned on and off at a very high speed (> 40 Gb/s). For this purpose, shorter carrier lifetime yields faster response, at the expense of higher dark current. Also, the depletion width should be minimized so the

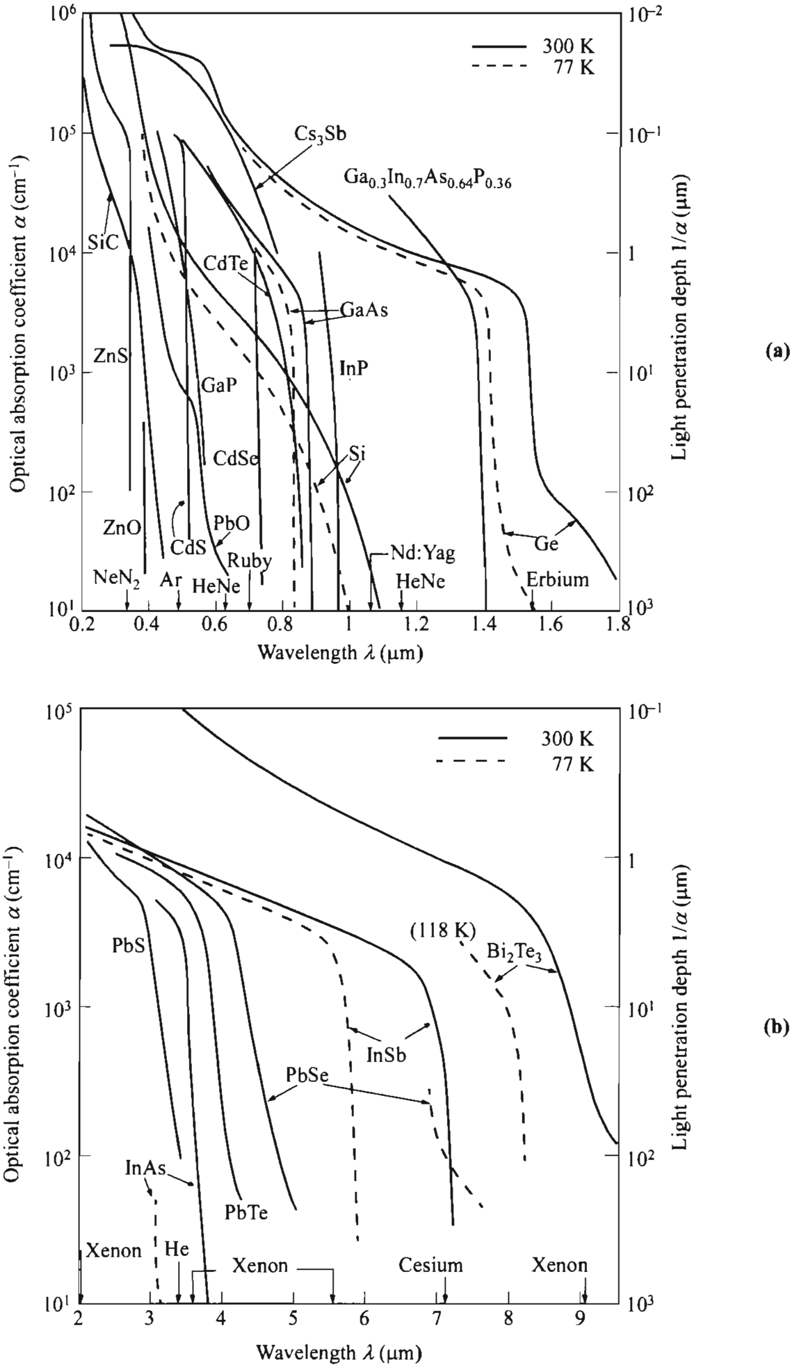


Fig. 1 Optical absorption coefficients for various photodetector materials, from (a) near optical to (b) infrared. Some laser emission wavelengths are indicated. (After Ref. 1.)

transit time can be shortened. On the other hand, the capacitance has to be kept low which means a larger depletion width. As seen, a trade-off has to be made for the overall optimization.

The signal of the photocurrent should be maximized for sensitivity. The basic metric is the quantum efficiency, defined as the number of carriers produced per photon, or

$$\eta = \frac{I_{ph}}{q\Phi} = \frac{I_{ph}}{q}\left(\frac{h\nu}{P_{\text{opt}}}\right) \tag{2}$$

where I_{ph} is the photocurrent, Φ is the photon flux ($= P_{\text{opt}}/h\nu$), and P_{opt} the optical power. The ideal quantum efficiency is unity. The reduction is due to current loss by recombination, incompletion absorption, reflection, etc. Another similar metric is the responsivity $\mathscr{R}$, using the optical power as the reference,

$$\mathscr{R} = \frac{I_{ph}}{P_{\text{opt}}} = \frac{\eta q}{h\nu} = \frac{\eta\lambda(\mu\text{m})}{1.24} \quad \text{A/W}. \tag{3}$$

To further improve the signal, some photodetectors have an internal gain mechanism. Comparison of the gains of common photodetectors are shown in Table 1. Gain as high as 10^6 can be achieved. Unfortunately, high gain also leads to higher noise which is the following topic.

Apart from a large signal, low noise is also important as it will ultimately determine the minimum detectable signal strength. That is why we often speak of the signal-to-noise ratio. There are many factors that contribute to noise. The dark current is the leakage current when the photodetector is under bias but not exposed to the light source. One limitation on the device operation is temperature so the thermal energy should be smaller than the photon energy ($kT < h\nu$). Another source of noise is from background radiation, such as black-body radiation from the detector housing at room temperature if not cooled. Internal device noise includes thermal noise (Johnson noise), which is related to the random thermal agitation of carriers in any resistive device. The shot noise is due to the discrete single events of the photoelectric effect, and the statistical fluctuations associated with them. This is especially important for low light intensity. The third is due to flicker noise, otherwise known as $1/f$

Table 1 Typical Values of Gain and Response Time for Common Photodetectors

Photodetector		Gain	Response time (s)
Photoconductor		1–10^6	10^{-8}–10^{-3}
Photodiodes	*p-n* junction	1	10^{-11}
	p-i-n junction	1	10^{-10}–10^{-8}
	Metal-semiconductor diode	1	10^{-11}
CCD		1	10^{-11}–10^{-4}*
Avalanche photodiode		10^2–10^4	10^{-10}
Phototransistor		$\approx 10^2$	10^{-6}

* Limited by charge transfer. Large integration time is an advantage for CCD for high sensitivity.

noise. This is due to random effects associated with surface traps and generally has $1/f$ characteristics that are more pronounced at low frequencies. The generation-recombination noise comes from the fluctuations of these generation and recombination events. Generation noise can originate from both optical and thermal processes.

Since all the noises are independent events, they can be added together as the total noise. A related figure-of-merit[2] is the noise-equivalent power (NEP) that corresponds to the incident rms optical power required to produce a signal-to-noise ratio of one in a 1-Hz bandwidth. To the first order, this is the minimum detectable light power. Finally, the detectivity D^* is defined as

$$D^* = \frac{\sqrt{AB}}{\text{NEP}} \qquad \text{cm-Hz}^{1/2}/W, \tag{4}$$

where A is the area and B is the bandwidth. This is also the signal-to-noise ratio when one watt of light power is incident on a detector of area 1 cm^2, and the noise is measured over a 1-Hz bandwidth. The parameter is normalized to the area since the device noise is generally proportional to the square root of area. The detectivity depends on the detector's sensitivity, spectral response, and noise. It is a function of wavelength, modulation frequency, and bandwidth, and is recommended to be expressed as $D^*(\lambda,f,B)$.

The last section of this chapter deals with solar cells which to some extent share some similarity with photodetectors in that they both convert light to electricity. The purpose of the solar cells, however, is for power generation from the sunlight, as opposed to detection of faint light. So one difference between them is the intensity of light involved. The second difference being that solar cells are power generators and as such no external bias is required, whereas photodetectors usually require some bias and the change of current is detected as signal.

13.2 PHOTOCONDUCTOR

A photoconductor consists simply of a slab of semiconductor, in bulk or thin-film form, with ohmic contacts affixed to the opposite ends (Fig. 2). When incident light falls on the surface of the photoconductor, carriers are generated either by band-to-band transitions (intrinsic) or by transitions involving forbidden-gap energy levels (extrinsic), resulting in an increase in conductivity. The processes of intrinsic and extrinsic photoexcitations of carriers are shown in Fig. 3.

For the intrinsic photoconductor, the conductivity is given by $\sigma = q(\mu_n n + \mu_p p)$, and the increase of conductivity under illumination is mainly due to the increase in the number of carriers. The wavelength cutoff is given by Eq. 1, where ΔE is the semiconductor bandgap E_g in this case. For shorter wavelengths, the incident radiation is absorbed by the semiconductor, and electron-hole pairs are generated. For the extrinsic photoconductor, photoexcitation occurs between a band edge and an impurity energy level in the energy gap.

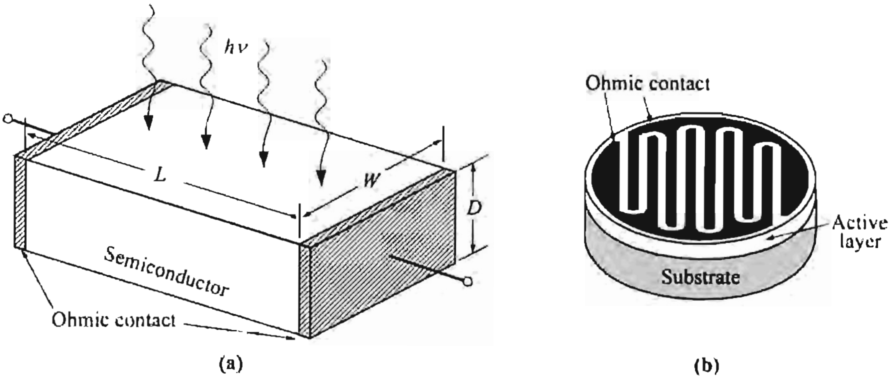


Fig. 2 (a) Schematic diagram of a photoconductor for analysis, which consists of a slab of semiconductor with two ohmic contacts. (b) Typical layout consists of interdigitated contacts with small gap.

The performance of a photodetector in general and a photoconductor in particular is measured in terms of three parameters: the quantum efficiency and gain, the response time, and the sensitivity (detectivity). First consider the principle of operation of a photoconductor under illumination (Fig. 2). Assuming a steady flow of photon flux impinging uniformly on the surface of a photoconductor with area $A = WL$, the total number of photons arriving at the surface is $(P_{opt}/h\nu)$ per unit time, where P_{opt} is the incident optical power and $h\nu$ is the photon energy. At steady state, the carrier generation rate G_e must be equal to the recombination rate. If the device thickness D is much larger than the light penetration depth $(1/\alpha)$ so that all light power is absorbed, the total steady-state generation and recombination rates of carriers per unit volume are

$$G_e = \frac{n}{\tau} = \frac{\eta(P_{opt}/h\nu)}{WLD} \tag{5}$$

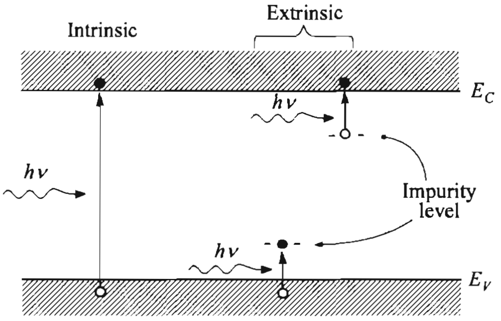


Fig. 3 Processes of intrinsic photoexcitation from band to band, and extrinsic photoexcitation between impurity level and band.

where τ is the carrier lifetime, η is the quantum efficiency (i.e., number of carriers generated per photon), and n is the excess carrier density. Since this concentration is much larger than the background doping level of the photoconductor, the steady-state concentration becomes

$$n = G_e \tau. \tag{6}$$

The carrier lifetime is related to the characteristics that if the light is taken off, the concentration would decay with time at a rate of

$$n(t) = n(0)\exp\left(\frac{-t}{\tau}\right). \tag{7}$$

For an intrinsic photoconductor, the photocurrent flowing between the electrodes is

$$I_p = \sigma \mathscr{E} WD = (\mu_n + \mu_p) nq \mathscr{E} WD \tag{8}$$

where $\mathscr{E}$ is the applied electric field inside the photoconductor, and $n = p$. Substituting n of Eq. 5 into Eq. 8 gives

$$I_p = q\left(\eta \frac{P_{\text{opt}}}{h\nu}\right)\frac{(\mu_n + \mu_p)\tau\mathscr{E}}{L}. \tag{9}$$

If we define the primary photocurrent as

$$I_{ph} \equiv q\left(\eta \frac{P_{\text{opt}}}{h\nu}\right), \tag{10}$$

the photocurrent gain G_a from Eq. 9 is

$$G_a = \frac{I_p}{I_{ph}} = \frac{(\mu_n + \mu_p)\tau\mathscr{E}}{L} = \tau\left(\frac{1}{t_{rn}} + \frac{1}{t_{rp}}\right) \tag{11}$$

where t_{rn} ($= L/\mu_n\mathscr{E}$) and t_{rp} ($= L/\mu_p\mathscr{E}$) are the electron and hole transit times across the electrodes. The gain depends upon the ratios of carrier lifetimes to the transit time and is a critical parameter in photoconductors. For high gain, the lifetime should be long, while the electrode spacing should be short and mobilities high. A typical gain of 1,000 is readily obtained, but higher gains up to 10^6 have been achieved (Table 1). On the other hand, the response time of a photoconductor is also determined by the lifetime. So there is a trade-off between gain and speed. A photoconductor generally has a response time much longer than that of a photodiode.

The high gain can be limited by the maximum field at breakdown. Another effect is due to the minority-carrier *sweep-out*.[3] Under a moderate field, the majority carriers (electrons) have a higher mobility and their transit time is shorter than the carrier lifetime. Meanwhile, the minority carriers (holes) are slower and their transit time is longer than the carrier lifetime. Under this condition, electrons are swept out of the detector quickly, but the holes demand charge neutrality and more electrons are supplied from the other electrode. Through this action, electrons are going through the detector many loops during the carrier lifetime, and this action is responsible for the gain. At very high fields, holes also move with a transit time shorter than the lifetime.

Under this condition, generation cannot keep up with the fast drift process, and the steady-state condition of Eq. 6 no longer holds. This condition results in the space-charge effect. At such a high field, the gain is degraded and it approaches unity again.

Next, consider an intensity-modulated optical signal given by

$$P(\omega) = P_{\text{opt}}[1 + m\exp(j\omega t)] \tag{12}$$

where P_{opt} is the average optical-signal power, m the modulation index, and ω the modulation frequency. The average current I_p resulting from the optical signal is given by Eq. 9. For the modulated optical signal, the rms optical power is $mP_{\text{opt}}/\sqrt{2}$ and the rms signal current can be written as[4]

$$i_p \approx \left(\frac{q\eta m P_{\text{opt}} G_a}{\sqrt{2}h\nu}\right)\frac{1}{\sqrt{1+\omega^2\tau^2}}. \tag{13}$$

At low frequencies, this reduces to Eq. 9. At high frequencies, the response is proportional to $1/f$.

Figure 4 shows an RF equivalent circuit for a photoconductor. The conductance G consists of the contributions from the dark current, the average signal current, and the background current. The thermal noise resulting from the conductance G is given by

$$\langle i_G^2 \rangle = 4kTGB \tag{14}$$

where B is the bandwidth. The generation-recombination noise (shot noise) is given by[5]

$$\langle i_{GR}^2 \rangle = \frac{4qI_p BG_a}{1+\omega^2\tau^2} \tag{15}$$

where I_p is the steady-state light-induced output current. The signal-to-noise ratio can be obtained from Eqs. 13–15:

$$\left.\frac{S}{N}\right|_{\text{power}} = \frac{i_p^2}{\langle i_{GR}^2\rangle + \langle i_G^2\rangle} = \frac{\eta m^2(P_{\text{opt}}/h\nu)}{8B}\left[1 + \frac{kT}{qG_a}(1+\omega^2\tau^2)\frac{G}{I_p}\right]^{-1}. \tag{16}$$

One can obtain the NEP (i.e., $mP_{\text{opt}}/\sqrt{2}$) from Eq. 16 by setting $S/N = 1$ and $B = 1$. For infrared detectors the most used figure of merit is the detectivity D^* which has been defined by Eq. 4.

The photoconductor is attractive for its simple structure, low cost, and rugged features. Extrinsic photoconductors can extend the long-wavelength limit without using materials of very narrow energy gap, and they are commonly used as infrared photodetectors. For mid-infrared to far-infrared and longer wavelengths, the photoconduc-

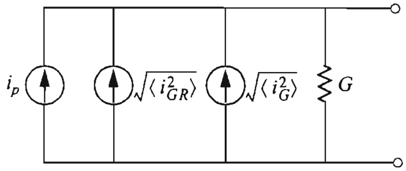

i_p
$\sqrt{\langle i_{GR}^2\rangle}$
$\sqrt{\langle i_G^2\rangle}$
G

Fig. 4 RF equivalent circuit of photoconductor. (After Ref. 4.)

tors are cooled to lower temperatures (such as 77 K and 4.2 K). The lower temperatures reduce thermal effects which cause thermal ionization and deplete the energy levels, and increase the gain and detection efficiency. Near 0.5 μm, a CdS photoconductor has high sensitivity, whereas at 10 μm a HgCdTe photoconductor is preferred.[6] In the wavelength range from 100 to 400 μm, a GaAs extrinsic photoconductor is a better choice because of its high detectivity.[7] This photoconductor has high dynamic range and can give comparable performance for high-level (strong light intensity) detection. For low-level detection at microwave frequencies, however, a photodiode will provide considerably more speed and higher signal-to-noise ratio. Thus photoconductors have limited use in high-frequency optical demodulators, such as in optical mixing. They have been, however, extensively used for infrared detection especially beyond a few microns of wavelength.

13.3 PHOTODIODES

13.3.1 General Consideration

A photodiode has a depleted semiconductor region with a high electric field that serves to separate photogenerated electron-hole pairs. For high-speed operation, the depletion region must be kept thin to reduce the transit time. On the other hand, to increase the quantum efficiency (the number of electron-hole pairs generated per incident photon), the depletion layer must be sufficiently thick to allow a large fraction of the incident light to be absorbed. Thus, there is a trade-off between the speed of response and quantum efficiency.

For the visible and near-infrared wavelength range, photodiodes are usually reverse-biased with moderate biasing voltages, because this reduces the carrier transit time and lowers the diode capacitance. The reverse voltage is, however, not large enough to cause avalanche multiplication or breakdown. This biasing condition is in contrast to avalanche photodiodes, where an internal current gain is obtained as a result of the impact ionization under avalanche breakdown conditions. All photodiodes, with the exception of the avalanche photodiode which is not included in this section, thus have a maximum gain of one (Table 1). The photodiode family includes the *p-i-n* photodiode, *p-n* photodiode, heterojunction photodiode, and metal-semiconductor (Schottky barrier) photodiode.

We shall now briefly consider the general characteristics of a photodiode: its quantum efficiency, response speed, and device noise.

Quantum Efficiency. As mentioned previously, quantum efficiency is the number of electron-hole pairs generated per incident photon (Eq. 2). A related figure of merit is the responsivity, which is the ratio of the photocurrent to the optical power (Eq. 3). Therefore, for a given quantum efficiency, the responsivity increases linearly with wavelength. For an ideal photodiode ($\eta = 1$), $\mathscr{R} = \lambda/1.24$ (A/W) where λ is expressed in microns.

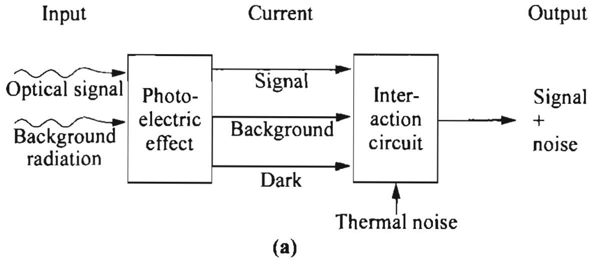
Since the optical absorption coefficient α is a strong function of the wavelength, for a given semiconductor the wavelength range in which appreciable photocurrent can be generated is limited. Since most photodiodes use band-to-band photoexcitation (except for photoexcitation over the barrier in metal-semiconductor photodiodes), the long-wavelength cutoff λ_c is established by the energy gap of the semiconductor, Eq. 1, for example about 1.7 μm for Ge and 1.1 μm for Si. For wavelengths longer than λ_c, the values of α are too small to give appreciable absorption. The short-wavelength cutoff of the photoresponse comes about because the values of α are very large ($\geq 10^5$ cm^{-1}), and the radiation is absorbed very near the surface where recombination is more likely. The photocarriers thus recombine before they are collected in the *p-n* junction. In the near-infrared region, silicon photodiodes with antireflection coating can reach 100% quantum efficiency near 0.8 to 0.9 μm. In the 1.0- to 1.6-μm region, Ge photodiodes, III-V ternary photodiodes (e.g., InGaAs), and III-V quaternary photodiodes (e.g., InGaAsP) have shown high quantum efficiencies. For longer wavelengths, photodiodes are cooled (e.g., 77 K) for high-efficiency operation.

Response Speed. The response speed is limited by a combination of three factors: (1) drift time in the depletion region, (2) diffusion of carriers, and (3) capacitance of the depletion region. Carriers generated outside the depletion region must diffuse to the junction resulting in considerable time delay. To minimize the diffusion effect, the junction should be formed very close to the surface. Most light will be absorbed when the depletion region is sufficiently wide (of the order of $1/\alpha$); with sufficient reverse bias the carriers will drift at their saturation velocities. The depletion layer must not be too wide, however, or transit-time effects will limit the frequency response. It also should not be too thin, or excessive capacitance C will result in a large R_LC time constant, where R_L is the load resistance. The optimum compromise occurs when the depletion layer is chosen so that the transit time is of the order of one-half the modulation period. For example, for a modulation frequency of 10 GHz, the optimum depletion layer thickness in Si (with a saturation velocity of 10^7 cm/s) is about 5 μm.

Device Noise. To study the noise properties in a photodiode, we will consider the generalized photodetection process shown in Fig. 5a. An optical signal and background radiation are absorbed by the photodiode, whereby electron-hole pairs are generated. These electrons and holes are then separated by the electric field and drift toward the opposite sides of the junction. In the process, a photocurrent is induced in the external load resistor. Since noise is frequency dependent, to determine the currents generated by this photoelectric process, we will consider an intensity-modulated optical signal given by Eq. 12. The average photocurrent I_P due to the optical signal is given by Eq. 10. For the modulated optical signal, the rms signal power is $mP_{\mathrm{opt}}/\sqrt{2}$, and the rms signal current is obtained from Eq. 13 with the gain set to unity,

$$i_p = \frac{q\eta m P_{\mathrm{opt}}}{\sqrt{2}h\nu}. \tag{17}$$

We designate the current resulting from the background radiation to be I_B, and the dark current due to thermal generation of electron-hole pairs in the depletion region


Input
Current
Output
Optical signal
Background radiation
Photo-electric effect
Signal
Background
Dark
Inter-action circuit
Signal + noise
Thermal noise
(a)

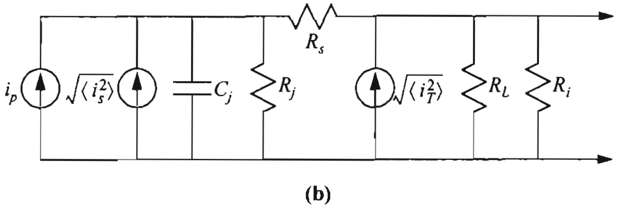

R_s
i_p
$\sqrt{\langle i_s^2 \rangle}$
C_j
R_j
$\sqrt{\langle i_T^2 \rangle}$
R_L
R_i
(b)

Fig. 5 Noise analysis of photodiode. (a) Photodetection process. (b) Equivalent circuit. (After Ref. 8.)

I_D. Because of the random generation of all these currents, they contribute to shot noise given by

$$\langle i_s^2 \rangle = 2q(I_P + I_B + I_D)B \tag{18}$$

where B is the bandwidth. The thermal noise is given by

$$\langle i_T^2 \rangle = \frac{4kTB}{R_{eq}} \tag{19}$$

where

$$\frac{1}{R_{eq}} = \frac{1}{R_j} + \frac{1}{R_L} + \frac{1}{R_i}. \tag{20}$$

The equivalent circuit of a photodiode is shown in Fig. 5b. The component C_j is the junction capacitance, R_j the junction resistance, and R_s the series resistance. The variable R_L is an external load resistor and R_i is the input resistance of the following amplifier.[9] All the resistances contribute additional thermal noise to the system. The series resistance R_s is usually much smaller than the other resistances and can be neglected.

For a 100% modulated signal ($m = 1$) with average power P_{opt}, the signal-to-noise ratio can be written as

$$\left.\frac{S}{N}\right|_{\text{power}} = \frac{i_p^2}{\langle i_s^2 \rangle + \langle i_T^2 \rangle} = \frac{(1/2)(q\eta P_{\text{opt}}/h\nu)^2}{2q(I_P + I_B + I_D)B + 4kTB/R_{eq}}. \tag{21}$$

From this equation, the minimum optical power required to obtain a given signal-to-noise ratio is (setting $I_P = 0$)

$$P_{\text{opt}}\big|_{\min} = \frac{2h\nu}{\eta}\sqrt{\frac{(S/N)I_{eq}B}{q}} \tag{22}$$

where

$$I_{eq} = I_B + I_D + \frac{2kT}{qR_{eq}}. \tag{23}$$

The noise-equivalent power (NEP) is given by ($S/N = 1$, $B = 1$ Hz)

$$\begin{aligned} \text{NEP} &= \text{rms optical power } P_{\text{opt}}\big|_{\min} \\ &= \left(\frac{h\nu}{\eta}\right)\sqrt{\frac{2I_{eq}}{q}} \quad \text{W/cm}^2\text{-Hz}^{1/2}. \end{aligned} \tag{24}$$

To improve the sensitivity of a photodiode, both η and R_{eq} should be increased while I_B and I_D should be minimized. The NEP decreases with R_{eq} until it saturates to a constant value limited by dark-current or background-current shot noise.

13.3.2 *p-i-n* and *p-n* Photodiodes

The *p-i-n* photodiode is a special case of the *p-n* junction photodiodes, and is one of the most-common photodetectors, because the depletion-region thickness (the intrinsic layer) can be tailored to optimize the quantum efficiency and frequency response. Figure 6 shows schematic representation of a *p-i-n* diode and its energy-band diagram under reverse-bias conditions, together with optical absorption characteristics. We shall discuss the operation of *p-i-n* photodiode in some detail with the help of Fig. 6. This discussion applies also to *p-n* junction photodiodes. Light absorption in the semiconductor produces electron-hole pairs. Pairs produced in the depletion region or within a diffusion length of it will eventually be separated by the electric field, leading to current flow in the external circuit as carriers drift across the depletion layer.

Quantum Efficiency. Under steady-state conditions the total photocurrent density through the reverse-biased depletion layer is given by[10]

$$J_{\text{tot}} = J_{dr} + J_{\text{diff}} \tag{25}$$

where J_{dr} is the drift current due to carriers generated within the depletion region and J_{diff} is the diffusion current due to carriers generated outside the depletion layer in the bulk of the semiconductor and diffusing into the reverse-biased junction. We shall now derive the total current under the assumptions that the thermal generation current can be neglected and that the surface p-layer is much thinner than $1/\alpha$. Referring to Fig. 6c, the electron-hole generation rate is given by

$$G_e(x) = \Phi_0\alpha\exp(-\alpha x) \tag{26}$$

where Φ_0 is the incident photon flux per unit area given by $P_{\text{opt}}(1 - R)/Ah\nu$, R is the reflection coefficient, and A is the device area. The drift current J_{dr} is thus given by

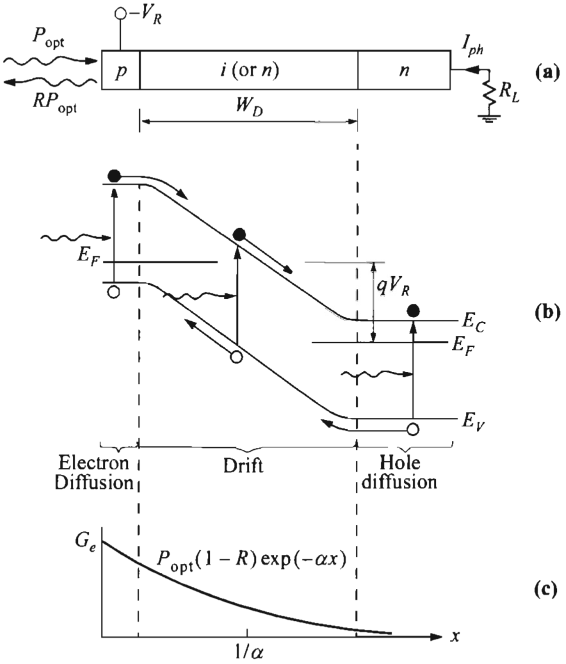

$-V_R$
P_{opt}
RP_{opt}
p
i (or n)
n
W_D
I_{ph}
R_L
(a)
E_F
qV_R
E_C
E_F
E_V
(b)
Electron Diffusion
Drift
Hole diffusion
G_e
$P_{opt}(1-R)\exp(-\alpha x)$
$1/\alpha$
x
(c)

Fig. 6 Operation of photodiode. (a) Cross-sectional view of *p-i-n* diode. (b) Energy-band diagram under reverse bias. (c) Carrier generation characteristics. (After Ref. 1.)

$$J_{dr} = -q\int_0^{W_D} G_e(x)dx = q\Phi_0[1-\exp(-\alpha W_D)] \tag{27}$$

where W_D is the depletion-layer width. Note that within the depletion region, a quantum efficiency of 100% has been assumed.

For $x > W_D$, the minority-carrier density (holes) in the bulk semiconductor is determined by the one-dimensional diffusion equation

$$D_p\frac{\partial^2 p_n}{\partial x^2} - \frac{p_n - p_{no}}{\tau_p} + G_e(x) = 0 \tag{28}$$

where D_p is the diffusion coefficient for holes, τ_p the lifetime of excess carriers, and p_{no} the equilibrium hole density. The solution of Eq. 28 under the boundary conditions $p_n = p_{no}$ for $x = \infty$, and $p_n = 0$ for $x = W_D$ is given by

$$p_n = p_{no} - [p_{no} + C_1\exp(-\alpha W_D)]\exp\left(\frac{W_D - x}{L_p}\right) + C_1\exp(-\alpha x) \tag{29}$$

with $L_p = \sqrt{D_p\tau_p}$ and

$$C_1 \equiv \left(\frac{\Phi_0}{D_p}\right)\frac{\alpha L_p^2}{1 - \alpha^2 L_p^2}. \tag{30}$$

The diffusion current density is given by

$$\begin{aligned} J_{\text{diff}} &= -qD_p \frac{\partial p_n}{\partial x}\bigg|_{x = W_D} \\ &= q\Phi_0 \frac{\alpha L_p}{1 + \alpha L_p}\exp(-\alpha W_D) + \frac{q p_{no} D_p}{L_p}. \end{aligned} \tag{31}$$

The total current density is the sum of I_{dr} inside the depletion region and I_{diff} outside the depletion, given by

$$J_{\text{tot}} = q\Phi_0\left[1 - \frac{\exp(-\alpha W_D)}{1 + \alpha L_p}\right] + \frac{q p_{no} D_p}{L_p}. \tag{32}$$

Under normal operating conditions, the dark-current term involving p_{no} is much smaller so that the total photocurrent is proportional to the photon flux. The quantum efficiency can be obtained from Eqs. 2 and 32,

$$\eta = \frac{AJ_{\text{tot}}/q}{P_{\text{opt}}/h\nu} = (1 - R)\left[1 - \frac{\exp(-\alpha W_D)}{1 + \alpha L_p}\right]. \tag{33}$$

Qualitatively, the quantum efficiency is reduced from unity due to refection R and light absorbed outside the depletion region. For high quantum efficiency, low reflection R and $\alpha W_D \gg 1$ is desirable. However, for $W_D \gg 1/\alpha$, the transit-time delay may be considerable. We consider next the transit-time effect.

Frequency Response. Since the carriers require a finite time to traverse the depletion layer, a phase difference between the photon flux and the photocurrent will appear when the incident light intensity is modulated rapidly. To obtain a quantitative result for this effect, the simplest case is shown in Fig. 7a where all the light is assume to be absorbed at the surface. The applied voltage is assumed to be high enough to deplete the intrinsic region and to ensure carrier saturation velocity υ_s. For a photon flux density given by $\Phi_1\exp(j\omega t)$ (photons/s-cm^2), the conduction current density J_{cond} at point x is found to be, assume $\eta = 100\%$,

$$J_{\text{cond}}(x) = q\Phi_1 \exp\left[j\omega\left(t - \frac{x}{\upsilon_s}\right)\right]. \tag{34}$$

The internal current is thus a function of time and distance. Since $\nabla \cdot J_{\text{tot}} = 0$, we can write the external total current as

$$J_{\text{tot}} = \frac{1}{W_D}\int_0^{W_D}\left(J_{\text{cond}} + \varepsilon_s \frac{\partial \mathscr{E}}{\partial t}\right)dx \tag{35}$$

where the second term in parentheses is the displacement current. Substituting Eq. 34 into Eq. 35 yields

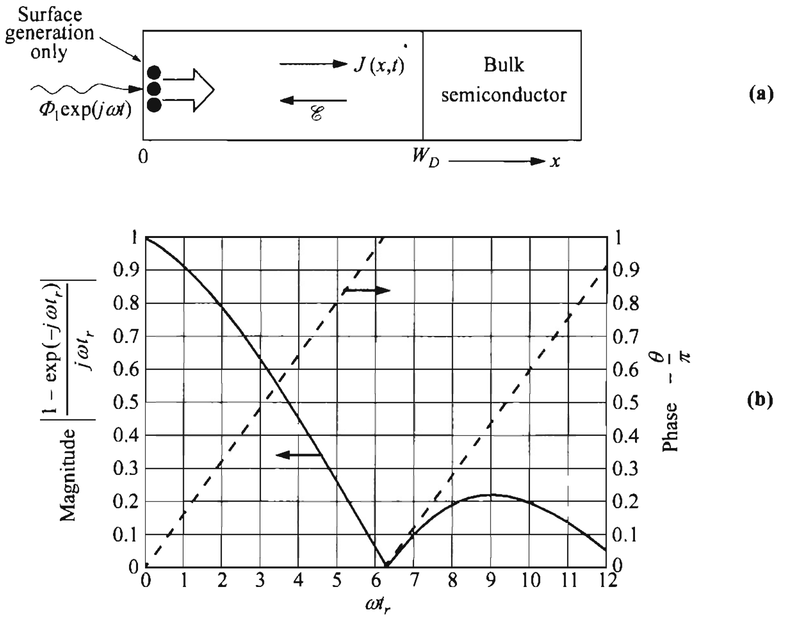


Fig. 7 (a) Geometry assumed for analysis of transit-time effect. Photoresponse (normalized magnitude and phase) vs. normalized modulation frequency of incident photon flux where $\theta = \omega t_r/2$. (After Ref. 10.)

$$J_{\text{tot}} = \left[\frac{j\omega\varepsilon_s V}{W_D} + q\Phi_1 \frac{1-\exp(-j\omega t_r)}{j\omega t_r}\right]\exp(j\omega t) \tag{36}$$

where V is the sum of applied voltage and the built-in potential, and $t_r = W_D/v_s$ is the transit time of carriers through the depletion region. From Eq. 36 the short-circuit current density ($V \approx 0$) is given by

$$J_{sc} = \frac{q\Phi_1[1-\exp(j\omega t_r)]}{j\omega t_r}\exp(j\omega t). \tag{37}$$

Figure 7b shows the transit-time effects at high frequencies where the amplitude and phase angle of the normalized current are plotted as a function of the normalized modulation frequency. Note that the magnitude of the ac photocurrent decreases rapidly with frequency when ωt_r exceeds unity. At $\omega t_r = 2.4$, the amplitude is reduced by $\sqrt{2}$ and is accompanied by a phase shift of 0.4π. The response time of the photodetector is thus limited by the carrier transit time through the depletion layer. A reasonable compromise between high-frequency response and high quantum efficiency is obtained for an absorption region of thickness $1/\alpha$ to $2/\alpha$ such that a reasonably large portion of the light is absorbed within the depletion region.

For the *p-i-n* photodiode, the thickness of the *i*-region is assumed equal to $1/\alpha$. The carrier transit time is the time required for carriers to drift through the *i*-region. From Eq. 37 the 3-dB frequency is given by ($\omega t_r = 2.4$)

$$f_{3\text{dB}} = \frac{2.4}{2\pi t_r} \approx \frac{0.4 v_s}{W_D} \approx 0.4 \alpha v_s . \tag{38}$$

Figure 8 shows the internal quantum efficiency, that is, $\eta/(1 - R)$ of the Si *p-i-n* photodiode as a function of the 3-dB frequency and the depletion width calculated from Eq. 38 and Fig. 1. The curves illustrate the trade-off between the response speed (3-dB frequency which is proportional to $1/W_D$) and quantum efficiency at various wavelengths, by adjusting the depletion width.

The constructions of some high-speed photodiodes are shown in Fig. 9, usually with an antireflection coating (not shown) to increase quantum efficiency. The *p-i-n* photodiode is shown in Fig. 9a. The thickness of the intrinsic region (or low *n*-type doping, ν-region, or low *p*-type doping, π-region) is optimized for the optical-signal wavelength and the modulation frequency. The *p-n* photodiode is a related structure where the *n*-type doping is high so that this layer is not fully depleted (Fig. 9b). At a wavelength close to the long-wavelength cutoff, the required absorption depth becomes very long (for $\alpha = 10\ \text{cm}^{-1}$, $1/\alpha = 1{,}000\ \mu\text{m}$). One option to compromise between quantum efficiency and response speed is to have the light incident from the side, parallel to the junction. This has the potential of reduced intrinsic layer thickness, shorter transit time, and thus higher speed, but at the expense of reduced quantum efficiency. The light can also be allowed to strike at an angle that creates multiple reflections inside the device, substantially increasing the effective absorp-

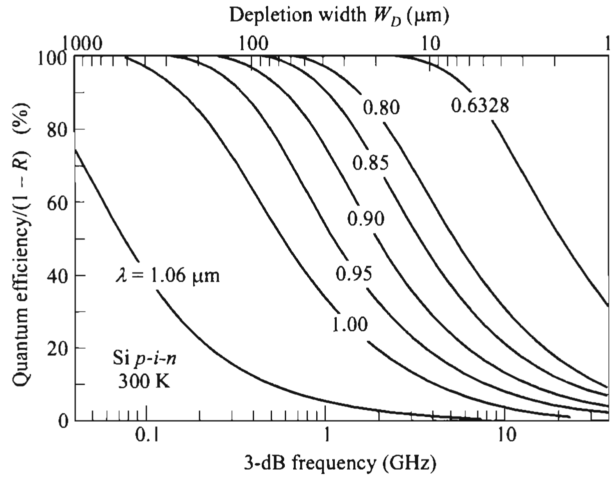


Fig. 8 Variation of quantum efficiency of Si *p-i-n* photodiode with depletion width and transit-time-limited 3-dB frequency for several wavelengths. Saturation velocity is 10^7 cm/s.

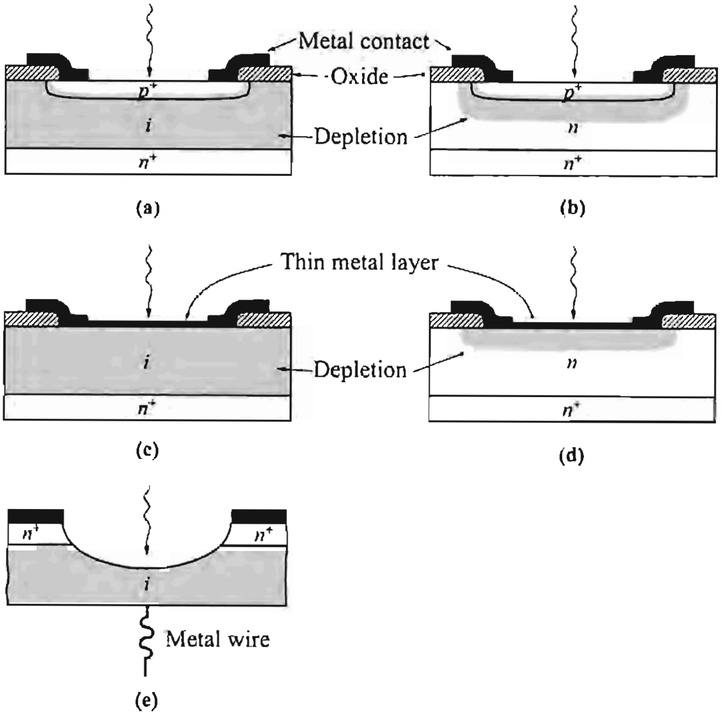


Fig. 9 Device configurations of some high-speed photodiodes. (a) *p-i-n* photodiode. (b) *p-n* photodiode. (c) Metal-*i-n* photodiode. (d) Metal-semiconductor photodiode. (e) Point-contact photodiode. (After Ref. 1.)

tion depth and at the same time keeping the carrier transit distance small.[11,12] The other three devices are metal-semiconductor photodiodes, to be considered later.

For the *p-n* photodiode, since the depletion-layer is thin, some portion of light could be absorbed outside the depletion region. This can lead to some disadvantages. First, the quantum efficiency is reduced. Light absorbed outside the depletion region by more than a diffusion length does not contribute to photocurrent at all, and that within a diffusion length the efficiency is also reduced. Second, the diffusion process is a slow one. The time required for carriers to diffuse a distance x is known to be

$$t = \frac{4x^2}{\pi^2 D_p}. \tag{39}$$

This is much slower than a drift process. The *p-n* photodiodes generally have a lower response speed than the *p-i-n* photodiodes. Finally, the neutral region contributes to series resistance which is a source of noise as discussed before.

13.3.3 Heterojunction Photodiode

A photodiode can be realized in a heterojunction which is formed between two semiconductors of different bandgaps (refer to Chapter 2). One major advantage of a heterojunction photodiode is that the quantum efficiency does not critically depend on the distance of the junction from the surface, because a large-bandgap material can be transparent and used as a window for the transmission of incoming optical power. In addition, the heterojunction can provide unique material combinations so that the quantum efficiency and response speed can be optimized for a given optical signal wavelength. Another advantage is a reduced dark current.

To obtain heterojunction with low leakage current, the lattice constants of the two semiconductors must be closely matched. Some examples of heterojunction photodiodes are given in Fig. 10, using an InP substrate and lattice matched to InGaAs (with $E_g \approx 0.73$ eV), and InAlAs. These structures have good performance at longer wavelengths (1 to 1.6 μm). This device is expected to have superior performance over the Ge photodiode because of its direct bandgap which gives rise to a larger absorption coefficient near the intrinsic absorption edge, so that a thinner depletion width can be used to give a higher response speed.[13] Another common system is AlGaAs on GaAs-substrate. These heterojunctions are important for photonic devices operated in the wavelength range of 0.65 to 0.85 μm.

13.3.4 Metal-Semiconductor Photodiode

A metal-semiconductor diode can be used as a high-efficiency photodetector.[14] The energy-band diagram and current transport in a metal-semiconductor diode have been considered extensively in Chapter 3. The photodiode can be operated in two modes, depending on the photon energy:

1. For $h\nu > E_g$, Fig. 11a, the radiation produces electron-hole pairs in the semiconductor, and the general characteristics of the photodiode are very similar to those

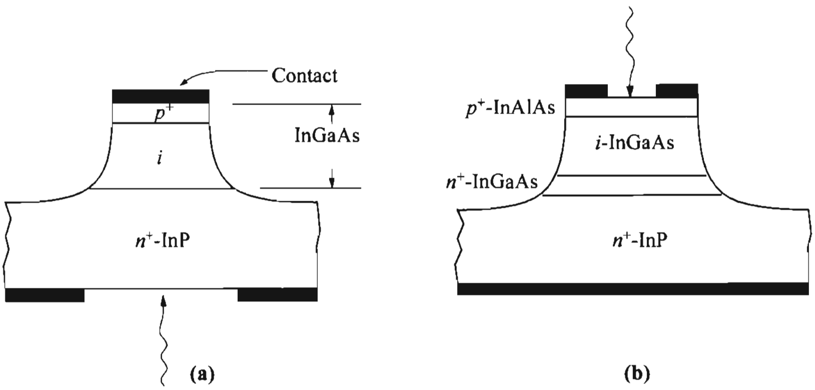


Fig. 10 Examples of heterojunction photodiodes on InP substrate (a) with substrate illumination and (b) with top illumination.

of a *p-i-n* photodiode. The quantum efficiency is given by an expression identical to Eq. 33.

2. For smaller photon energy (longer wavelength) $q\phi_B < h\nu < E_g$, Fig. 11b, the photoexcited electrons in the metal can surmount the barrier and be collected by the semiconductor. This process is called internal photoemission and has been used extensively to determine the Schottky-barrier height and to study the hot-electron transport in metal films.[15]

For the first process $h\nu > E_g$ and with high reverse bias at breakdown, the diode can be operated as an avalanche photodiode. This will be included in the discussion in the next section—avalanche photodiode.

For internal photoemission, the photon is absorbed in the metal layer and a carrier is excited to a higher energy. These hot carriers have momentum in random directions, and those having excess energy larger than the barrier height and momentum

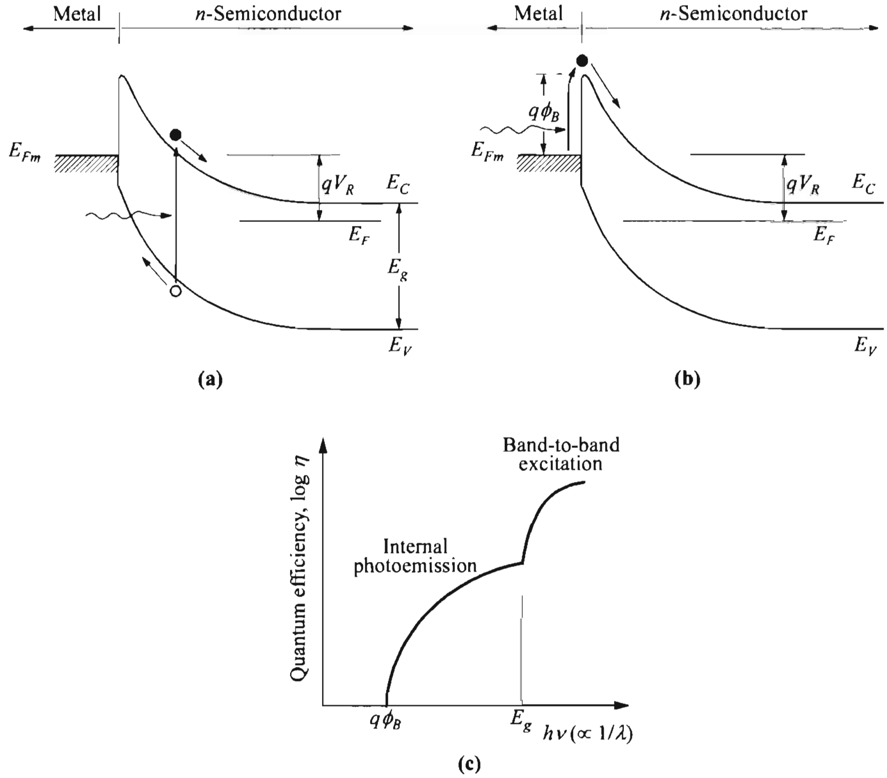

Metal
n-Semiconductor
E_{Fm}
qV_R
E_C
E_F
E_g
E_V
(a)
Metal
n-Semiconductor
$q\phi_B$
E_{Fm}
qV_R
E_C
E_F
E_V
(b)
Quantum efficiency, log η
Band-to-band excitation
Internal photoemission
$q\phi_B$
E_g
$h\nu(\propto 1/\lambda)$
(c)

Fig. 11 (a) Band-to-band excitation of electron-hole pair ($h\nu > E_g$). (b) Internal photoemission of excited electrons from metal to semiconductor ($E_g > h\nu > q\phi_B$). (c) Quantum efficiency as a function of wavelength showing both processes.

toward the semiconductor contribute to the photocurrent. The internal photoemission process is energy dependent, and the quantum efficiency is given by

$$\eta = C_F \frac{(h\nu - q\phi_B)^2}{h\nu}, \tag{40}$$

where C_F is the Fowler emission coefficient. This phenomenon is often applied to measure the barrier height. When a Schottky-barrier diode is scanned with light of variable wavelength, Fig. 11c shows that the quantum efficiency has a threshold of $q\phi_B$, and it increases with the photon energy. When the photon energy reaches the energy-gap value, the quantum efficiency jumps to a much higher value. In practical applications, the internal photoemission has typical quantum efficiencies of only less than 1%.

A typical configuration is illustrated in Fig. 9c. To avoid large reflection and absorption losses when the diode is illuminated through the metal contact, the metal film must be very thin, ≈ 10 nm, and an antireflection coating must be used. By using a low-doping *i*-layer, a metal-*i*-*n* photodiode similar to a *p*-*i*-*n* diode can be made. This structure is advantageous mainly for band-to-band excitation. A special metal-semiconductor diode is the point-contact photodiode shown in Fig. 9e.[16] The active volume is very small, and as a result both the drift time and capacitance are small. It is thus suitable for very-high modulation frequencies.

For detectors with internal photoemission, it is more efficient to direct the incoming light through the substrate. Since the barrier height is always smaller than the energy gap, light with $q\phi_B < h\nu < E_g$ is not absorbed in the semiconductor, and intensity is not reduced at the metal/semiconductor interface. The metal layer, in this case, can be thicker for easier thickness control and to minimize series resistance. For Si devices, options are available using silicides in place of the metal. A silicide usually has a more reproducible interface since it is formed by reacting metal with Si so that the new interface is never exposed. Common silicides used for this purpose are PtSi, Pd_2Si, and IrSi. Another advantage of a Schottky-barrier diode is that high-temperature processing for diffusion or implantation anneal is not required.

Metal-semiconductor photodiodes are particularly useful in the visible and ultraviolet regions. In these regions the absorption coefficients α in most of the common semiconductors are very high, of the order of 10^5 cm^{-1} or higher, corresponding to an effective absorption length of $1/\alpha = 0.1$ µm or less. It is possible to choose a proper metal and antireflection coating so that a large fraction of the incident radiation will be absorbed near the surface of the semiconductor.

The dark current in a Schottky diode is known to be due to thermionic emission of majority carriers and it does not suffer from charge storage of minority diffusion current, which limits the speed capability in a *p*-*n* photodiode. Ultrafast Schottky-barrier photodiodes operating beyond 100 GHz have been reported. The main advantages of the Schottky-barrier photodiode are high speed and long-wavelength detection capability without having to use a semiconductor with a small energy gap.

13.4 AVALANCHE PHOTODIODE

Avalanche photodiodes (APDs) are operated at high reverse-bias voltages where avalanche multiplication takes place.[17] The multiplication gives rise to internal current gain. The current gain-bandwidth product of an APD can be higher than 300 GHz, so the device can respond to light modulated at microwave frequencies. For APDs, the criteria with respect to quantum efficiency and response speed are similar to those for nonavalanching photodiodes. However, the high gain comes with the price of noise, so we must consider the noise properties as well as the avalanche gains.

13.4.1 Avalanche Gain

The avalanche gain, also called multiplication factor, has been considered in Chapter 2. The low-frequency avalanche gain for electrons is given by

$$M = \left\{1 - \int_0^{W_D} \alpha_n \exp\left[-\int_x^{W_D} (\alpha_n - \alpha_p) dx'\right] dx\right\}^{-1} \tag{41}$$

where W_D is the depletion-layer width and α_n and α_p are the electron and hole ionization rates, respectively. For position-independent ionization coefficients, as in a *p-i-n* diode, the multiplication of electrons injected into the high-field region at $x = 0$ is

$$M = \frac{(1 - \alpha_p/\alpha_n)\exp[\alpha_n W_D(1 - \alpha_p/\alpha_n)]}{1 - (\alpha_p/\alpha_n)\exp[\alpha_n W_D(1 - \alpha_p/\alpha_n)]}. \tag{42}$$

For equal ionization coefficients ($\alpha = \alpha_n = \alpha_p$), the multiplication takes the simple form

$$M = \frac{1}{1 - \alpha W_D}. \tag{43}$$

The breakdown voltage corresponds to the situation where $\alpha W_D = 1$.

In a practical device, the maximum achievable dc multiplication at high light intensities is limited by the series resistance and the space-charge effect. These factors can be combined into one effective series resistance R_s. The multiplication for photogenerated carriers can be described by an empirical relationship as[18]

$$M_{ph} = \frac{I - I_{MD}}{I_P - I_D} = \left[1 - \left(\frac{V_R - IR_s}{V_B}\right)^n\right]^{-1} \tag{44}$$

where I is the total multiplied current, I_P is the primary (unmultiplied) photocurrent, and I_D and I_{MD} are the primary and multiplied dark currents, respectively. V_R is the reverse-bias voltage, V_B is the breakdown voltage, and the exponent n is a constant depending on the semiconductor material, doping profile, and radiation wavelength. For high light intensity ($I_P \gg I_D$) and $IR_s \ll V_B$, the maximum value of the photomultiplication is given by

$$M_{ph}\Big|_{\max} \approx \frac{I}{I_P} = \left[1 - \left(\frac{V_R - IR_s}{V_B}\right)^n\right]^{-1}\Bigg|_{V_R \to V_B} \approx \frac{V_B}{nIR_s} \tag{45}$$

or

$$M_{ph}\big|_{\max} = \sqrt{\frac{V_B}{nI_PR_s}}. \tag{46}$$

When the photocurrent is smaller than the dark current, the maximum multiplication is limited by the dark current and is given by an expression similar to Eq. 46, except I_P is replaced by I_D. Thus, it is important that the dark current be as low as possible so that it will not limit either the $(M_{ph})_{\max}$ or the minimum detectable power.

The regenerative avalanche process results in the presence of a large number of carriers in the high-field region long after the primary electrons have traversed through that region. The higher the avalanche gain (or multiplication) is, the longer it takes to build up the avalanche process, and after the light is taken off, the longer the avalanche process persists. This implies a behavior that is set by a gain-bandwidth ($M{\cdot}B$) product. Figure 12 shows the calculated bandwidth for an idealized *p-i-n* avalanche photodiode with an avalanche region of uniform electric field. The 3-dB bandwidth B, normalized to $2\pi\tau_{av}$, is plotted as a function of the low-frequency gain M, with the ratio of the ionization coefficients as a parameter. The dashed curve is for $M = \alpha_n/\alpha_p$. Below this curve where $M > \alpha_n/\alpha_p$, the curves are almost straight lines, indicating a constant gain-bandwidth product. In this regime, the frequency dependence of gain is given by[19]

$$M_f(\omega) = \frac{M}{\sqrt{1 + [\omega M N(\alpha_p/\alpha_n)\tau_{av}]^2}}. \tag{47}$$

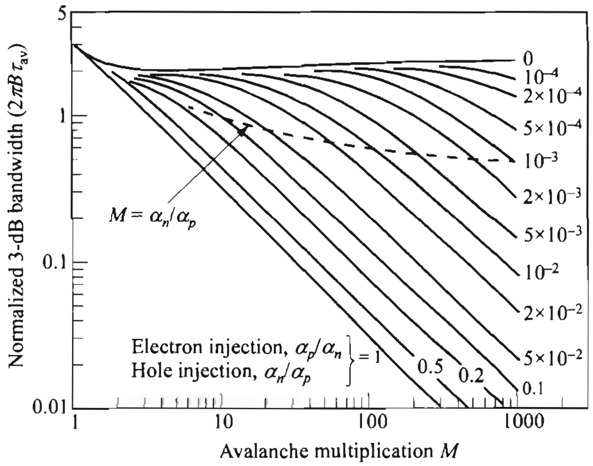


Fig. 12 Theoretical 3-dB bandwidth B (times $2\pi\tau_{av}$) of APD as a function of low-frequency multiplication M, for various values of α_p/α_n for electron injection (or α_n/α_p for hole injection). (After Ref. 19.)

Here N is a function of the ratio α_p/α_n. It has a value of 1/3 when $\alpha_p/\alpha_n = 1$ and a value of 2 when $\alpha_p/\alpha_n = 10^{-3}$. The average transit time τ_{av} is $(t_{rn} + t_{rp})/2$ where t_{rn} is the electron transit time, equal to W_D/v_{sn} and v_{sn} is the electron saturation velocity. A similar expression is found for hole transit time t_{rp}. From Eq. 47, the bandwidth B is obtained by setting the second term in the denominator equal to one, and the gain-bandwidth product is given by

$$M \cdot B = \frac{1}{2\pi N(\alpha_p/\alpha_n)\tau_{\text{av}}}. \tag{48}$$

For the special case of equal ionization coefficients and large gain, the gain-bandwidth product is found to be $M{\cdot}B = 3/2\pi\tau_{\text{av}}$. To obtain large gain-bandwidth products, v_{sn} and v_{sp} should be large, and α_p/α_n and W_D should be small. Above this dashed curve where $M < \alpha_n/\alpha_p$, the bandwidth is largely determined by the transit time of the carriers and is essentially independent of gain.

13.4.2 Avalanche-Multiplication Noise

The avalanche process is statistical in nature because every electron-hole pair generated at a given distance in the depletion region is independent and does not experience the same multiplication. Since the avalanche gain fluctuates, the mean-square value of the gain $\langle M^2\rangle$ is greater than the square of the mean $\langle M\rangle^2$. The excess noise can be characterized by a noise factor

$$F(M) \equiv \frac{\langle M^2\rangle}{\langle M\rangle^2} = \frac{\langle M^2\rangle}{M^2}. \tag{49}$$

The noise factor $F(M)$ is a measure of the increase in the shot noise compared to an ideal noiseless multiplier and it strongly depends on the ratio of the ionization coefficients α_p/α_n and on the low-frequency multiplication factor M. We will show that the noise factor $F(M)$ is always equal to or greater than unity and increases monotonically with multiplication except for a noiseless multiplication process. When $\alpha_n = \alpha_p$, on the average, only three carriers, the primary and its secondary hole and electron, are present in the multiplying region for every incident photocarrier. A fluctuation that changes the number of carriers by one represents a large percentage change, and the noise factor will be large. On the other hand, if one of the ionization coefficients approaches zero (e.g., $\alpha_p \rightarrow 0$), carriers of the order M are present in the multiplying region for every incident photocarrier; a fluctuation of one carrier is a relatively insignificant perturbation. Thus, the noise factor is expected to be small if the difference between α_n and α_p is large.

For electron injection alone, the noise factor can be written as[20]

$$\begin{aligned} F &= M\left[1 - (1-k)\left(\frac{M-1}{M}\right)^2\right] \\ &\approx kM + \left(2 - \frac{1}{M}\right)(1-k) \end{aligned} \tag{50}$$

where $k \equiv \alpha_p/\alpha_n$ is assumed to be constant throughout the avalanche region. For hole injection alone, the foregoing expression still applies if k is replaced by $k' \equiv \alpha_n/\alpha_p$. For the two special cases; $\alpha_p = \alpha_n$ (i.e., $k = 1$), Eq. 50 gives $F = M$, and if $\alpha_p \to 0$ (i.e., $k = 0$), we have $F = 2$ (at large M). The noise factor for various values of multiplication and ratios of ionization coefficients is shown in Fig. 13. We can see that a smaller value k for electron injection or a small value of k' for hole injection is desirable to minimize excess noise.

Figure 14 shows some experimental results obtained from a Si avalanche photodiode with a 0.1-μA primary injection current measured at 600 kHz. The upper values (open circles) represent the noise for a hole primary photocurrent, which results from short wavelength radiation (see the inset). The lower values (closed circles) represent the noise for an electron primary photocurrent. The noise factor for electron injection is considerably lower than that for hole injection, because α_n is much larger than α_p in silicon. The good agreement between theory and experiments can be seen.

The results shown in Fig. 13 can be used for the *p-i-n* APD and the lo-hi-lo type APD (discussed later), which has a uniform electric field in the avalanche region. For a general APD with nonuniform electric field, the ionization coefficients must be weighted accordingly: k is replaced by k_{eff}, and k' is replaced by k'_{eff} in Eq. 50, where[22]

$$k_{\text{eff}} = \int_0^{W_D} \alpha_p(x) M^2(x) dx \bigg/ \int_0^{W_D} \alpha_n(x) M^2(x) dx , \tag{51}$$

$$k'_{\text{eff}} = k_{\text{eff}} \left[\int_0^{W_D} \alpha_p(x) M(x) dx \bigg/ \int_0^{W_D} \alpha_n(x) M(x) dx \right]^{-2} . \tag{52}$$

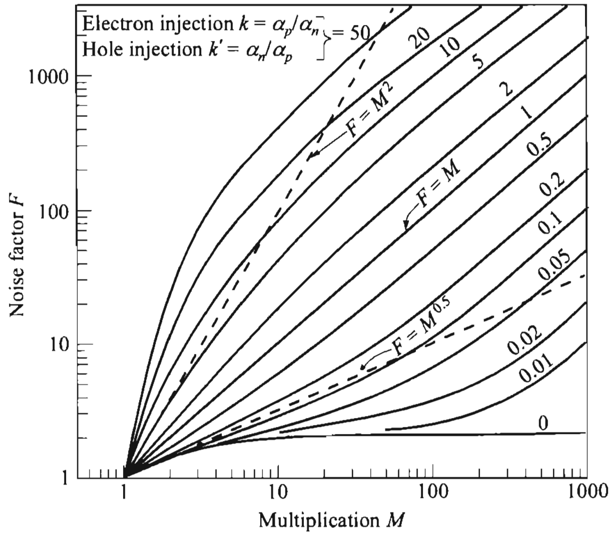


Fig. 13 Theoretical noise factor vs. multiplication, for different ratios of electron and hole ionization coefficients. (After Ref. 20.)

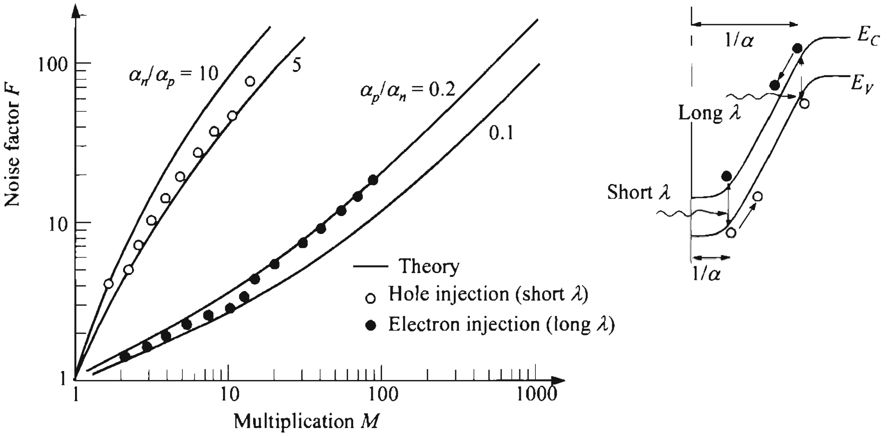


Fig. 14 Experimental results of noise factors for a silicon APD with 0.1-μA primary current of two wavelengths. Inset shows energy-band diagram with an electron or hole primary current, depending on the wavelength of the incident light. (After Ref. 21.)

Additional noise is introduced when light is absorbed on both sides of the junction so that both electrons and holes are injected into the avalanche region. For example, for $k_{\text{eff}} = 0.005$ and $M = 10$, the noise factor increases from about 2 for pure electron injection to 20 for 10% electron injection.[23] Therefore, to achieve low noise and wide bandwidth in an APD, the ionization coefficients of the carriers should be as different as possible and the avalanche process should be initiated by the carrier species with the higher ionization rate. The other species with lower ionization rate should be kept to a minimum as primary photogenerated current for noise consideration, and so it is advantageous to avoid light absorption inside the high-field avalanche region, as will be shown later.

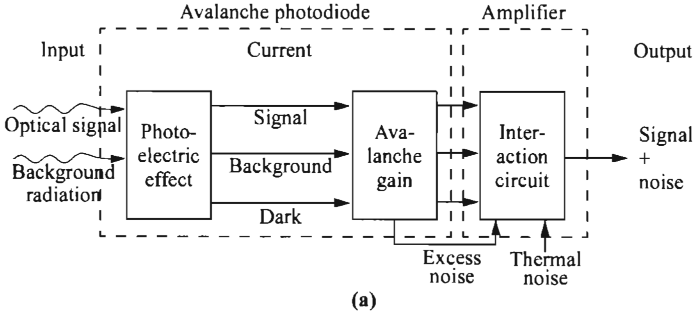
13.4.3 Signal-to-Noise Ratio

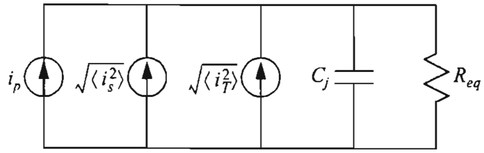
The photodetection process and equivalent circuit for an avalanche photodiode are shown schematically in Fig. 15a. The current gain mechanism multiplies the signal current, background current, and dark current indiscriminately. The multiplied rms signal photocurrent is identical to that of Eq. 17 except for the addition of the multiplication factor or avalanche gain M,

$$i_p = \frac{q\eta m P_{\text{opt}} M}{\sqrt{2}h\nu}. \tag{53}$$

The other elements of the equivalent circuit in Fig. 15b are the same as for the *p-i-n* photodiode. The mean-square shot-noise current after multiplication is given by

$$\begin{aligned} \langle i_s^2 \rangle &= 2q(I_P + I_B + I_D)\langle M^2 \rangle B \\ &= 2q(I_P + I_B + I_D)M^2F(M)B. \end{aligned} \tag{54}$$


Avalanche photodiode
Amplifier
Input
Current
Output
Optical signal
Background radiation
Photo-electric effect
Signal
Background
Dark
Ava-lanche gain
Inter-action circuit
Signal + noise
Excess noise
Thermal noise
(a)


i_p
$\sqrt{\langle i_s^2 \rangle}$
$\sqrt{\langle i_T^2 \rangle}$
C_j
R_{eq}

(b)

Fig. 15 (a) Photodetection process in avalanche photodiode. (b) Equivalent circuit. (After Ref. 8.)

The thermal noise is the same as for the *p-i-n* photodiode and is given by Eq. 19.

For a 100% modulated signal with average power P_{opt}, the signal-to-noise power ratio for the APD is then

$$\frac{S}{N} = \frac{(1/2)(q\eta P_{\mathrm{opt}}/h\nu)^2}{2q(I_P + I_B + I_D)F(M)B + 4kTB/(R_{eq}M^2)} . \tag{55}$$

From Eq. 55 we see that the avalanche gain can increase the signal-to-noise ratio by reducing the importance of the last term in the denominator. The *S/N* ratio increases with *M* until *F(M)* also becomes large. Thus, there is an optimum value of *M* which produces the maximum *S/N* ratio for a given optical power. This optimum multiplication is obtained when the first term in the denominator is approximately equal to the second term. The optimum multiplication M_{opt} is found by setting $d(S/N)/dM = 0$. Substituting this M_{opt} into Eq. 55, we obtain a maximum signal-to-noise ratio under the large signal photocurrent condition:[24]

$$\left.\frac{S}{N}\right|_{\max} \propto \frac{\eta}{\sqrt{k}} . \tag{56}$$

Therefore to maximize *S/N*, we should maximize $\eta/\sqrt{k}$.

Equation 55 can be solved for the minimum optical power P_{opt} required to produce a given *S/N* with avalanche gain. This power has the same expression as Eq. 22 except now

$$I_{eq} \equiv (I_B + I_D)F(M) + \frac{2kT}{qR_{eq}M^2} . \tag{57}$$

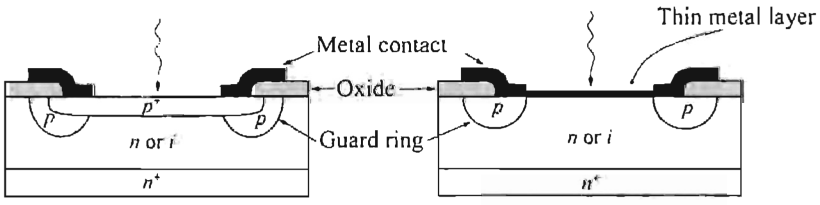
The noise equivalent power NEP is the same as Eq. 24. This NEP is improved through the reduction of I_{eq} by the gain M. Since avalanche gain can substantially reduce the NEP, the APDs can have a significant advantage over unity-gain photodiodes.

13.4.4 Device Performance

An avalanche photodiode requires the avalanche multiplication to be spatially uniform over the entire area of the diode.[25] Microplasmas, that is, small areas in which the breakdown voltage is less than that of the junction as a whole, must be eliminated. The probability of microplasmas occurring in the active area is minimized by using low dislocation materials and by designing the active area to be no larger than necessary to accommodate the incident light beam, generally from a few μm to 100 μm in diameter. The excessive leakage current along the junction edges due to the junction curvature effect or high-field concentration is eliminated by using a guard-ring or surface-beveled structure.[26]

Figure 16 shows some basic APD device configurations. Their main difference from a regular photodiode is the addition of guard rings at the junction perimeter to control leakage current under high bias. The guard-ring profile must have a low impurity gradient with a sufficiently large radius of curvature such that the guard-ring junction will not breakdown before the central p^+-n (or p-i-n) junction does. For the metal-semiconductor APD, a guard ring must also be used to eliminate high electric field concentration at the periphery of the contact (Fig. 16b). A mesa or beveled structure can have a low surface field across the junction and uniform avalanche breakdown can occur inside the device (not shown). This is more common for compound-semiconductor devices due to their inferior planar technology. To detect wavelength near the intrinsic absorption edge, a side-illuminated APD can be used to improve both the quantum efficiency and the signal-to-noise ratio.

Avalanche photodiodes have been made in various semiconductors including Ge, Si, and III-V compounds and their alloys. The key factors in selecting a particular semiconductor include the quantum efficiency at a particular optical wavelength, the response speed, and the noise. We shall now consider some representative device performances.



(a) (b)

Fig. 16 Basic device configurations of avalanche photodiodes. (a) p-n or p-i-n structure. (b) Metal-semiconductor structure. Note guard rings at the junction perimeters.

Germanium APDs are useful in the wavelength range from 1 to 1.6 μm because of high quantum efficiency. Since the ionization coefficients of electrons and holes are comparable in Ge, the noise factor is close to $F = M$ (Eq. 50), and the mean-square shot noise current varies as M^3 (Eq. 54).[1] For intermediate gains $M < 30$, the signal power increases as M^2 and the noise power as M^3. This behavior is in good agreement with the theoretical prediction. The highest S/N ratio (≈ 40 dB) is obtained at $M \approx 10$, that is, where the noise contribution from the diode is about equal to the receiver noise. At higher values of M, the S/N ratio decreases because avalanche noise increases faster than the multiplied signal.

Silicon APDs are particularly useful in the wavelength range from 0.6 to 1.0 μm, where nearly 100% quantum efficiency has been obtained from devices having antireflection coatings. The hole-to-electron ionization coefficient ratio ($k = \alpha_p/\alpha_n$) in silicon is a strong function of the electric field; it varies from about 0.1 at 3×10^5 V/cm to 0.5 at 6×10^5 V/cm. Therefore, to minimize noise, the electric field at avalanche breakdown should be low and the ionization multiplication should be initiated by electrons.

Some idealized doping profiles are shown in Fig. 17 where there are two regions of different field strengths. The wide and low-field region serves as light absorption and the narrow and high-field region for avalanche multiplication. These are called reach-through structures because the electric field extends all the way from the n^+-layer to the p^+-layer (fully depleted).[27] The doping profile of a p^+-π-p-π-n^+ structure is shown in Fig. 17a. This profile is similar to that of a lo-hi-lo IMPATT diode (see Chapter 9). In the lower-field drift region for absorption, the carriers can travel at their saturation velocity (10^7 cm/s for $\mathscr{E}_d > 10^4$ V/cm). In the high-field avalanche region, the maximum field $\mathscr{E}_m$ can be adjusted by adjusting the thickness b. The breakdown condition can be written as[28]

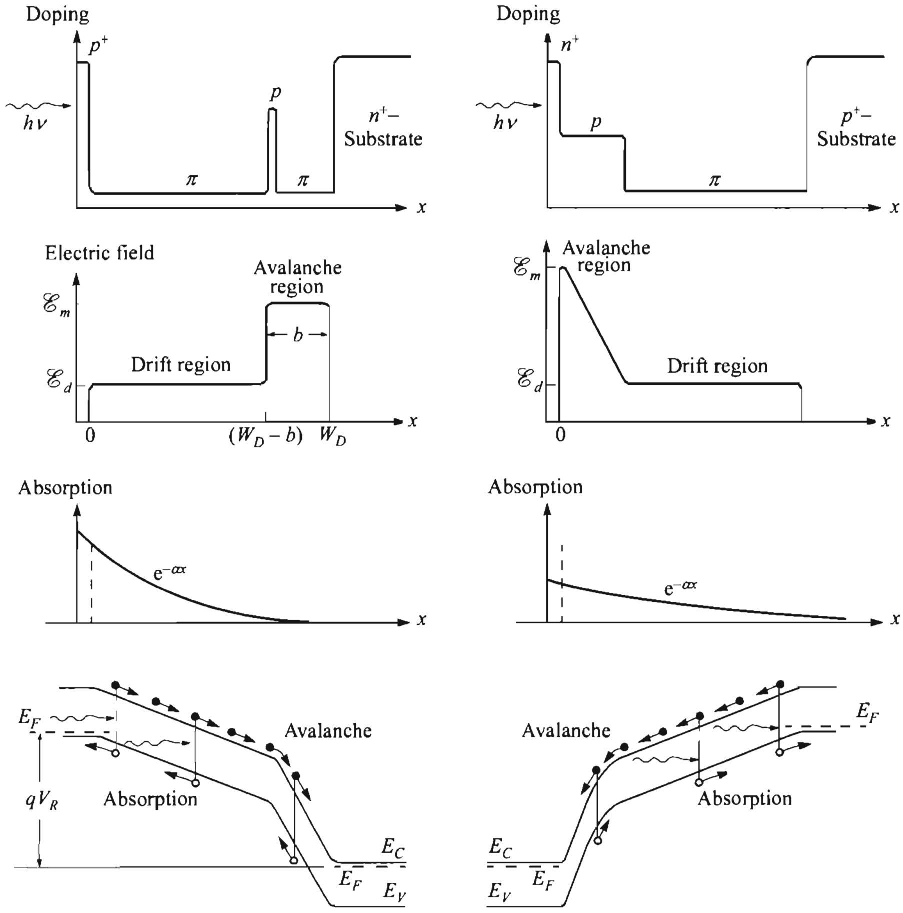
$$\alpha_n b = \frac{\ln(k)}{k-1} \qquad k \equiv \frac{\alpha_p}{\alpha_n}, \tag{58}$$

and the breakdown voltage is given by

$$V_B \approx \mathscr{E}_m b + \mathscr{E}_d(W_D - b). \tag{59}$$

For a given wavelength, we can choose a W_D (e.g., $W_D = 1/\alpha$) and then independently adjust b to optimize device performances. Most of the light should be absorbed in the π-region ($W_D - b$), and electrons enter the avalanche region to initiate the multiplication process. The p^+-π-p-π-n^+ device is expected to have high quantum efficiency, high response speed, and good signal-to-noise ratio.

In practice, it may be difficult to form the narrow p-region and the n^+-p-π-p^+ device (Fig. 17b) can be an option. This doping profile is identical to a hi-lo IMPATT structure. This device configuration is more amenable to fabrication on large-diameter silicon wafers, with good control of the doping profile through ion implantation and diffusion.[29] The quantum efficiency is near 100% at about 0.8-μm wavelength for a device having an antireflection coating. Because of the slight mixture of holes in initiating the multiplication process, the noise factor is higher than for the structure shown in Fig. 17a.


Doping
p^+
$h\nu$
p
π
π
n^+–
Substrate
x
Electric field
$\mathscr{E}_m$
$\mathscr{E}_d$
Avalanche region
b
Drift region
0
$(W_D - b)$
W_D
x
Absorption
$e^{-\alpha x}$
x
E_F
qV_R
Absorption
Avalanche
E_C
E_F
E_V
Doping
n^+
$h\nu$
p
π
p^+–
Substrate
x
$\mathscr{E}_m$
Avalanche region
$\mathscr{E}_d$
Drift region
0
x
Absorption
$e^{-\alpha x}$
x
Avalanche
E_F
Absorption
E_C
E_F
E_V

(a) **(b)**

Fig. 17 Reach-through avalanche photodiodes with doping profile, field distribution, absorption of light, and energy-band diagram, showing how electrons initiate the multiplication process. (a) Lo-hi-lo APD. (b) Hi-lo APD.

Metal-semiconductor (Schottky-barrier) APDs are useful in the visible and ultraviolet range. However, they are much less common compared to junctions formed by doping because of higher inherent leakage in Schottky barriers under high bias. The basic characteristics of Schottky-barrier APDs are similar to those of *p*-*n* junction APDs. Schottky-barrier APD on 0.5-Ω-cm *n*-type silicon substrate with a thin PtSi film (≈ 10 nm) and a diffused guard ring, as shown in Fig. 16b, with ideal reverse saturation current had been obtained. For the Schottky-barrier APD, avalanche multiplication can amplify the peak value of fast photocurrent pulses by factors up to 35.[30] Noise measurements for avalanche multiplication in PtSi-Si APDs show that the noise of the multiplied photocurrent increases approximately as M^3 for light in the visible range. As the wavelength decreases, the electron primary injection photocurrent becomes dominant and the noise decreases, in agreement with noise theory.

Schottky-barrier APDs with *n*-type silicon substrates promise to be particularly useful as high-speed photodetectors for ultraviolet light. Ultraviolet light that is transmitted through the thin metal electrodes is absorbed within the first 10 nm of the silicon. The carrier multiplication is then mainly initiated by electrons, resulting in low noise and high gain-bandwidth product. Amplification of high-speed photocurrent pulses is also possible. Bear in mind that photoexcitation over the barrier can occur, which extends the wavelength range beyond that of the energy bandgap (see Fig. 11b).

Heterojunction avalanche photodiodes, especially in III-V alloys, have many advantages as alternatives to Ge and Si devices. By adjusting the alloy composition, the wavelength response of the device can be tuned. Because of the high absorption coefficients of the direct-bandgap III-V alloys, the quantum efficiency can be high, even if a narrow depletion width is used to provide high-speed response. In addition, the heterostructure window layer (larger bandgap for the surface layer) can be grown so that high-speed performance is obtained and the surface recombination loss of photogenerated carriers can be minimized.

Heterostructure APDs have been made in various alloy systems such as AlGaAs/GaAs, AlGaSb/GaSb, InGaAs/InP, and InGaAsP/InP. These structures have shown improvements in speed and quantum efficiency over that obtained for Ge and Si. Extensive studies are continuing in this field to understand the material quality, absorption coefficients, and reliability. Many heterojunction APDs are made using III-V semiconductors grown on GaAs- or InP-substrates. The ternary or quaternary compound with a closely matched lattice parameter is then grown epitaxially on the substrate (e.g., by liquid- or vapor-phase epitaxy or molecular-beam epitaxy). The composition of the alloys, doping concentrations, and layer thicknesses are adjusted to optimize device performance.

A common configuration is an AlGaAs/GaAs heterojunction. The top AlGaAs layer serves as a window for the transmission of 0.5- to 0.9-µm incident light. The ionization-coefficient ratio k $(= \alpha_p/\alpha_n)$ is not preferable in $\langle 100 \rangle$-oriented GaAs (= 0.83). For $\langle 111 \rangle$-oriented GaAs, the hole ionization rate is much larger than that for electrons (refer to Chapter 1). To minimize the avalanche noise, we should use $\langle 111 \rangle$-oriented GaAs employing holes to initiate the multiplication process.

One of the main advantages of a heterojunction APD is the use of higher bandgap material in the multiplication region while keeping the lower bandgap material for light absorption. Since the breakdown voltage V_B is expected to vary as $E_g^{3/2}$, the dark current due to tunneling and microplasmas are greatly suppressed. This effect also prevents edge breakdown in the APD structures. This approach has been called *separate absorption and multiplication*.

Figure 18 shows an example of a heterojunction APD having separate absorption and multiplication regions, based on the InGaAs/InP system.[31] The p^+-n junction is formed in InP (multiplication region) and due to its larger E_g, light is not absorbed. The InGaAs layer grown on the n-InP is used as a light absorption region whose

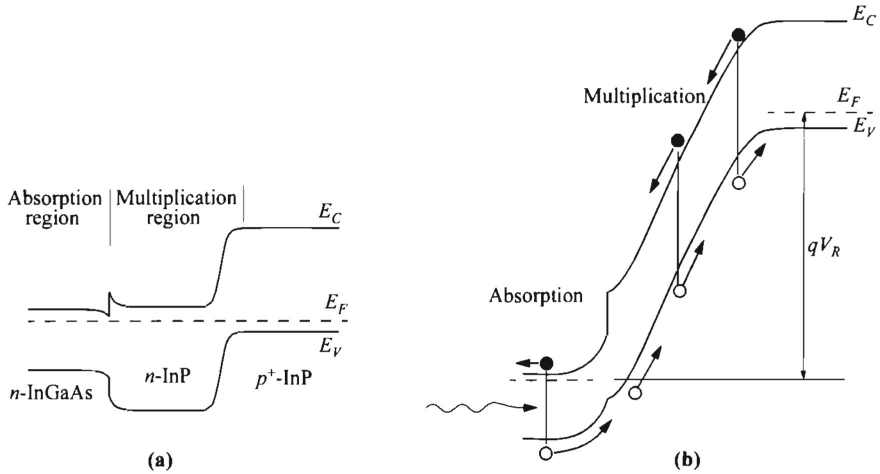


Fig. 18 Energy-band diagrams of InGaAs/InP heterojunction APD (a) at thermal equilibrium. (b) with avalanche multiplication.

lower E_g is required by the wavelength of interest. Because the ionization rate of holes is larger by two to three times than that of electrons in InP ($k' = 0.4$), the avalanche process should be initiated by holes. The dopings and thicknesses of n-InP and n-InGaAs layers are designed so that under avalanching conditions, the n-InP layer is fully depleted (Fig. 18b). The composition of InP near the heterojunction also has to be graded to avoid a barrier for holes in the valence band ΔE_V that would accumulate holes. The device has a quantum efficiency of 40% at 1.3 μm and 50% at 1.6 μm. Its noise factor is 3 dB lower than that of Ge APD operated at 1.15 μm.

An added advantage of a heterojunction APD is that if the multiplication region is made sufficiently thin, the noise can be further reduced. Qualitatively, impact ionization requires some minimum distance, often called *dead space*, for carriers to gather sufficient energy from the field. Longer multiplication region allows more multiplication and larger gains which, in turn, produces larger statistical fluctuations. This ultimately leads to more noise. Such phenomenon is demonstrated in Fig. 19 where it is apparent that the distribution at high gain is tightened when the multiplication region is reduced from 1 μm to 0.1 μm, while the average gain is the same for both (≈ 20). The noise factor is reduced from 6.9 to 4, accordingly.[17]

Since noise is an important issue for APDs, some material properties had been exploited to improve the ratio of the ionization rates. Studies on $Al_xGa_{1-x}Sb$ junctions show that when the spin-orbit split Δ of the valence bands approaches the bandgap (Fig. 20, inset), the value of k' can become very small.[32] Figure 20 shows a pronounced reduction of k' at $\Delta/E_g \approx 1$. Values of k' smaller than 0.04 have been obtained, corresponding to noise factors less than 5 at $M = 100$. Such phenomenon can be observed in other materials such as InGaAsSb and HgCdTe.

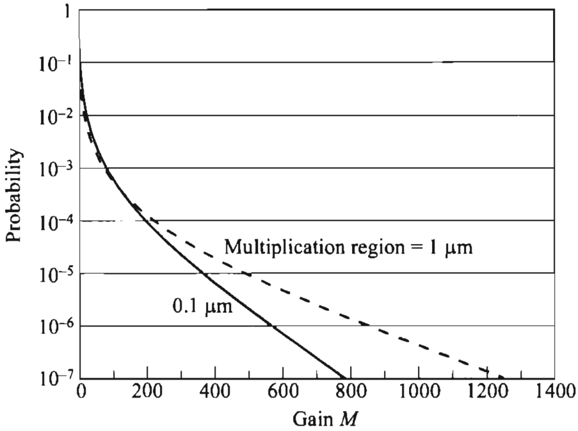


Fig. 19 Gain distribution of InAlAs APDs having multiplication regions of 1 μm and 0.1 μm. The average gain for both is the same (≈ 20). (After Ref. 17.)

13.5 PHOTOTRANSISTOR

A phototransistor can have high gain through the internal bipolar-transistor action. On the other hand, the fabrication of a phototransistor is more complicated than that of a photodiode, and the inherent larger area degrades its high-frequency performance. Compared to an avalanche photodiode, it eliminates the high voltage required and high noise associated with avalanche, yet provides reasonable photocurrent gain.

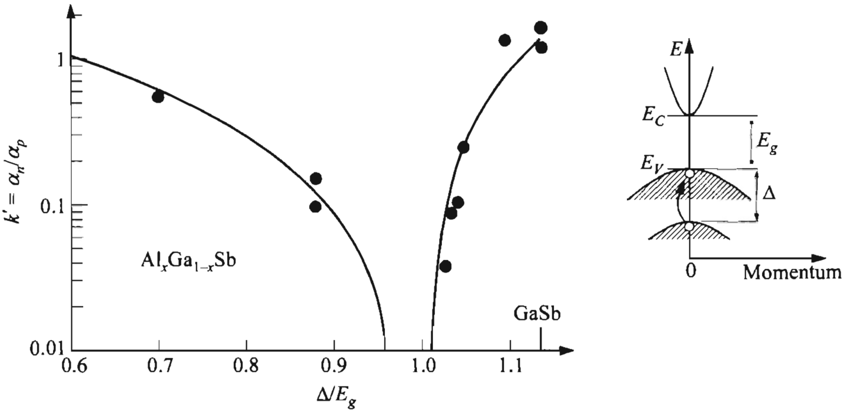


Fig. 20 Ratio of ionization rates in $Al_xGa_{1-x}Sb$ vs. Δ/E_g, where Δ is the spin-orbit splitting of the valence bands shown in the inset. (After Ref. 32.)

A bipolar phototransistor is shown in Fig. 21, together with its circuit model. It differs from a conventional bipolar transistor by having a large base-collector junction as the light-collecting element, represented by a parallel combination of a diode and a capacitor. This device is particularly useful in opto-isolator applications because it offers a high current-transfer ratio, i.e., the ratio of output photodetector current to the input light-source (LED or laser) current, of the order of 50% or more, as compared to a typical photodiode with a current-transfer ratio of 0.2%.

The phototransistor is biased in the active regime. For a floating base, that simply means positive bias to the collector with respect to the emitter for an *n-p-n* structure. The energy-band diagram illustrating the response to light is shown in Fig. 21c. Photogenerated holes, in the base/collector depletion region and within a distance of the diffusion length, flow to the energy maximum and are trapped in the base. This accumulation of holes or positive charges lowers the base energy (raises the potential) and allows a large flow of electrons from the emitter to the collector. The result of a much larger electron current caused by a small hole current is the consequence of emitter injection efficiency γ and is the dominant gain mechanism that is common for both the bipolar transistor and the phototransistor, provided that the electron transit time through the base is much shorter than the minority-carrier lifetime. The photogenerated electrons, depending on the location of origin, can flow to the emitter or to the collector. Strictly speaking, they can reduce the emitter current or enhance the col-

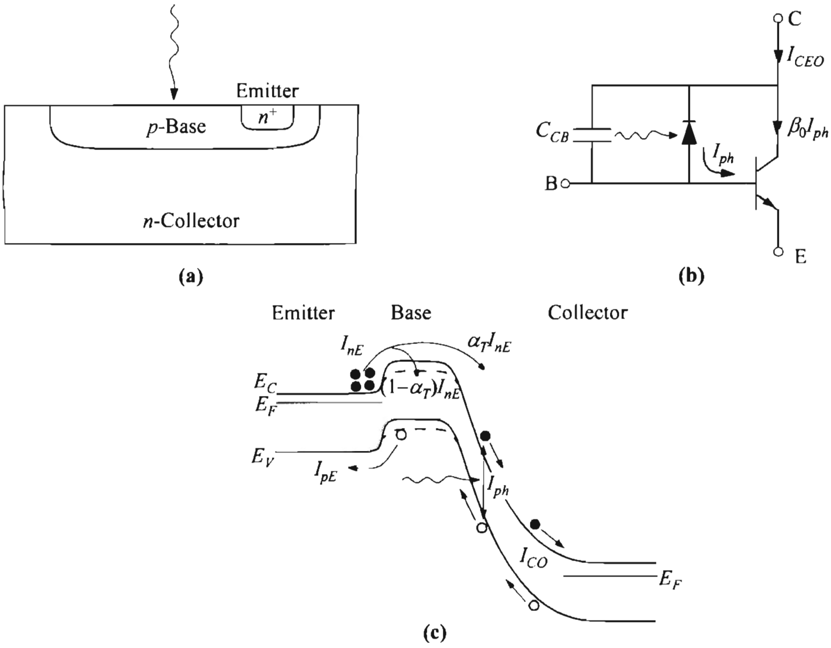


Fig. 21 (a) Schematic structure of phototransistor. (b) Equivalent circuit. (c) Energy-band diagram under bias showing different current components. Dashed lines indicate shift of base potential (open base) under illumination.

lector current, but only by a very small amount since the gain is large and the total collector current or emitter current is much larger than the photocurrent. For simplicity, the following analysis assumes that light is absorbed near the base-collector junction as shown in Fig. 21c.

From this figure, and using the conventional bipolar transistor parameters summarized in Table 2 of Chapter 5, the total collector current is given by

$$I_C = I_{ph} + I_{CO} + \alpha_T I_{nE} \tag{60}$$

where I_{ph} is the photocurrent, I_{CO} is the reverse leakage current of the collector-base junction, and α_T is the base transport factor. Since the base is open, the net base current is zero and

$$I_{pE} + (1 - \alpha_T) I_{nE} = I_{ph} + I_{CO} . \tag{61}$$

From Eqs. 60 and 61, and the definition of emitter injection efficiency γ,

$$I_{nE} = \gamma I_E , \tag{62}$$

it can be shown that

$$I_{CEO} = (I_{ph} + I_{CO})(\beta_0 + 1) \approx \beta_0 I_{ph} . \tag{63}$$

The *I-V* characteristics of a phototransistor under different light intensities are similar to that of the bipolar transistor, except the base incremental current is replaced with increasing light intensities (Fig. 8b of Chapter 5). Equation 63 indicates a photocurrent gain of $(\beta_0 + 1)$. Unfortunately the dark current is also amplified by the same factor. In practical homojunction phototransistors, gains vary from 50 to a few hundred. For heterojunction phototransistors, gains up to 10k can be obtained. One drawback of the phototransistor is that the gain is not constant with light intensity since the latter affects the base potential, so linearity is compromised.

The speed of a phototransistor is limited by the charging times of the emitter and the collector, indicated by

$$\begin{aligned} \tau &= \tau_E + \tau_C \\ &= \beta_0 \left[\frac{kT}{qI_{CEO}} (C_{EB} + C_{CB}) + R_L C_{CB} \right] , \end{aligned} \tag{64}$$

where C_{EB}, C_{CB} are the emitter-base and collector-base capacitances respectively, and R_L is the load resistance. In practical homojunction devices, the response time is relatively long, usually in the range 1–10 μs, limiting the operational frequency to ≈ 200 kHz. The frequency of heterojunction phototransistors can go beyond 2 GHz. Several observations can be made from Eq. 64. First, the speed goes up when the light signal (or I_{CEO}) is larger. In applications where speed is critical, the device is made with a base contact, and an applied dc bias increases the dc collector current. The trade-off is reduced photocurrent gain. Second, the speed is inversely proportional to the gain. For this reason, a gain-bandwidth product is a better measure of the performance.

The noise equivalent power is given by an expression similar to Eq. 24 in which[33]

$$I_{eq} = I_{CEO}\left(1 + \frac{2h_{fe}^2}{h_{FE}}\right) \tag{65}$$

where h_{fe} is the incremental common-emitter current gain. Therefore, there is a tradeoff between low noise and high gain.

By adding a second bipolar transistor, a Darlington phototransistor (or photo-Darlington) with even higher transfer ratio can be formed (Fig. 22). One of the transistors serves as a phototransistor, with the emitter current fed to the base of the other transistor, which acts as an additional amplifier. For the first order, the gain becomes β_0^2. The frequency response of this structures is limited by the large base-collector capacitance and is reduced further by the gain of the detector due to feedback. For comparison, typical response time for a photodiode is of the order of 0.01 μs, while it is about 5 μs for a phototransistor, and 50 μs for a Darlington phototransistor.

The heterojunction phototransistor, whose emitter has a larger energy gap than the base, can have advantages similar to that of a regular heterojunction bipolar transistor. Heterostructures studied include AlGaAs/GaAs, InGaAs/InP, and CdS/Si. The emitter with a wider energy gap has higher injection efficiency, leading to higher gain, and it allows the base to be more heavily doped for lower base resistance. It can also be transparent to the incoming light, so that light is more efficiently absorbed in the base and the collector. A double-heterojunction phototransistor has an additional heterojunction at the collector-base junction.[34] The device shows high blocking voltage and high gain for both bias polarities, and linear current-voltage characteristics through the zero bias point. A bilateral gain of higher than 3,000 has been obtained.

13.6 CHARGE-COUPLED DEVICE (CCD)

The charge-couple device (CCD) can be used either as an image sensor or as a shift register. In fact, when used in imaging array systems such as a camera or video recorder, they are functioning as both. As a photodetector, it has also been called charge-coupled image sensor or charge-transfer image sensor. As a signal shifter, it is

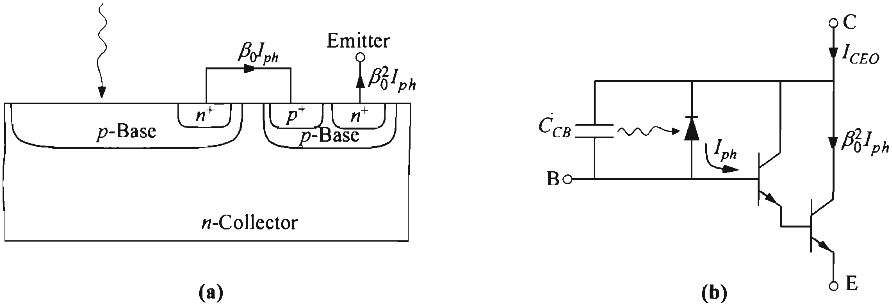
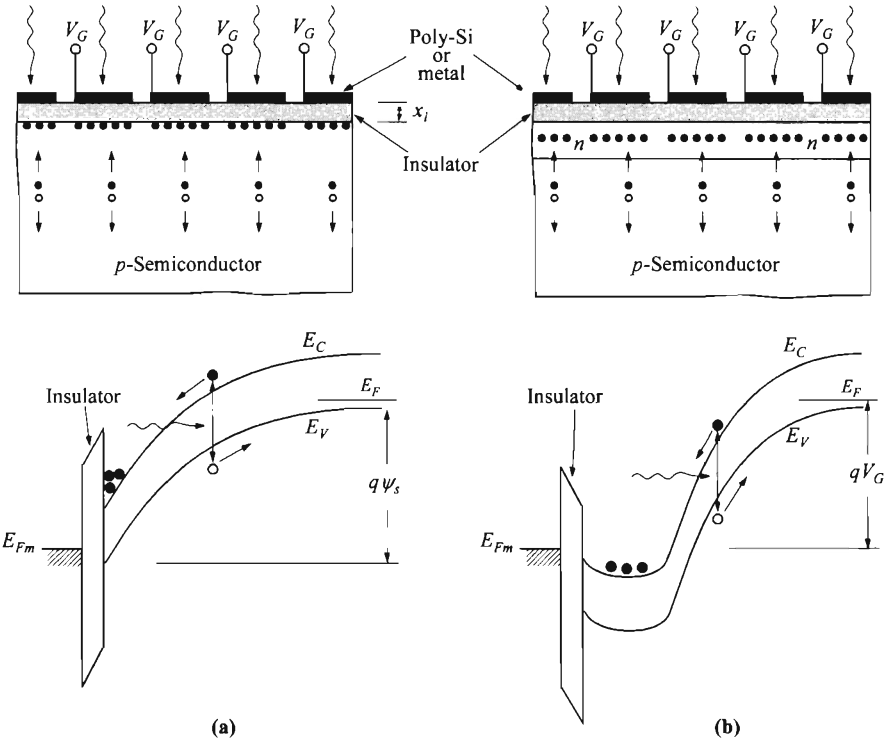


Fig. 22 (a) Schematic structure of Darlington phototransistor and (b) its equivalent circuit.

also called a charge-transfer device. When the concept of CCD was introduced mainly as shift register by Boyle and Smith in 1970, the possibility of using it as an imaging device was briefly mentioned in their seminal paper.[35] The idea sparked intensive research activities immediately. The CCD as a linear scanning system was first demonstrated in 1970.[36,37] This was later extended to an area scanning system in 1972.[38] To extend the wavelength beyond that detectable by Si devices, compound semiconductors started to be examined in 1973.[39,40] Since the 1970s, the CCD has developed into a mature technology for commercial imaging products.

13.6.1 CCD Image Sensor

The structures of the surface-channel CCD image sensor are similar to those of the CCD shift register, with the exception that the gates are semitransparent to let light pass through (Fig. 23a). Common materials for the gates are metal, polysilicon, and silicide. Alternatively, the CCD can be illuminated from the back of the substrate to avoid light absorption by the gate. In this configuration, the semiconductor has to be thinned down so that most of the light can be absorbed within the depletion region at



(a) (b)

Fig. 23 Structures and energy-band diagrams of (a) surface-channel CCD and (b) buried-channel CCD. For p-type substrate, a positive gate bias is applied to drive the semiconductor into deep depletion under nonequilibrium condition.

the top surface, and that spatial resolution is not lost since each pixel typically is less than 10 μm on the side. Unlike other photodetectors, CCD image sensors must be spaced close to one another in a chain. This is due to the unique feature that they can function as shift registers to transport the signals. Figure 23b shows a buried-channel CCD (BCCD) with a layer of opposite type at the surface. This thin layer (≈ 0.2–0.3 μm) is fully depleted and the accumulated photogenerated charge is kept away from the surface. This structure has the advantages of higher transfer efficiency and lower dark current, from reduced surface recombination. The penalty is smaller charge capacity, by a factor of 2–3 compared to the surface-channel CCD (SCCD). The most-common semiconductor used for CCD is Si, although other materials such as HgCdTe and InSb have been explored.

The CCD photodetector is unique in that there is no external dc photocurrent during light exposure. The photogenerated carriers are integrated during light exposure, and the signal is stored in the form of a charge packet, to be transported and detected later. This is somewhat similar to a photodiode (*p-i-n* or Schottky) operated under an open-circuit condition. Since each CCD is basically an MIS (metal-insulator-semiconductor) capacitor, it has to be operated in a nonequilibrium condition under a large gate pulse. If the semiconductor is allowed to recover from deep depletion, collection of the photogenerated charge would not be efficient, as discussed later.

For the sake of simplicity, we limit our discussion to surface-channel devices. The energy-band diagram immediately after applying a large gate pulse is shown in Fig. 23a. The gate bias has the polarity that drives the semiconductor to deep depletion. For an empty well, the gate voltage and the surface potential ψ_s under deep depletion are related by

$$V_G - V_{FB} = V_i + \psi_s = \frac{qN_A W_D}{C_i} + \psi_s \tag{66}$$

where V_i is the voltage across the insulator, C_i is the insulator capacitance (ε_i/x_i), and

$$\psi_s = \frac{qN_A W_D^2}{2\varepsilon_s}. \tag{67}$$

The depletion width W_D is larger than the maximum depletion width under equilibrium. Eliminating W_D from Eqs. 66 and 67 yields a relationship between the gate voltage and the surface potential,

$$V_G - V_{FB} = \psi_s + \frac{\sqrt{2\varepsilon_s qN_A \psi_s}}{C_i}. \tag{68}$$

The large surface potential creates a potential well for the photogenerated electrons while the photogenerated holes diffuse to the substrate. Similar to that in a photodiode, the internal quantum efficiency η within the depletion width W_D is close to 100%, and the overall η with frontal illumination is given by

$$\eta = 1 - \frac{\exp(-\alpha W_D)}{1 + \alpha L_n} \tag{69}$$

where L_n is the electron diffusion length. The total signal charge density Q_{sig} is thus proportional to the light intensity and the total exposure time,

$$Q_{\text{sig}} = -q\Phi\int \eta dt \tag{70}$$

where Φ is the photon flux density.

As the electrons start to accumulate at the semiconductor surface, the field across the insulator starts to increase, and the surface potential and the depletion width begin to shrink. With a signal charge packet present at the semiconductor surface, the surface field and the oxide field become

$$\mathscr{E}_s = \frac{qN_A W_D + Q_{\text{sig}}}{\varepsilon_s} = \sqrt{\frac{2qN_A\psi_s}{\varepsilon_s}}, \tag{71a}$$

$$\mathscr{E}_i = \frac{qN_A W_D - Q_{\text{sig}}}{\varepsilon_i} = \frac{V_i}{x_i}. \tag{71b}$$

Equation 66 now becomes

$$V_G - V_{FB} = \frac{\sqrt{2\varepsilon_s qN_A\psi_s} - Q_{\text{sig}}}{C_i} + \psi_s. \tag{72}$$

Equation 72 can be solved for ψ_s, resulting in

$$\psi_s = V_G - V_{FB} + \frac{qN_A\varepsilon_s}{C_i^2} + \frac{Q_{\text{sig}}}{C_i} - \frac{1}{C_i}\sqrt{2qN_A\varepsilon_s\left(V_G - V_{FB} + \frac{Q_{\text{sig}}}{C_i}\right) + \left(\frac{qN_A\varepsilon_s}{C_i}\right)^2}. \tag{73}$$

So for a given gate voltage, ψ_s decreases essentially linearly as the stored charge increases. It can be shown that the maximum signal that can be collected is

$$Q_{max} \approx C_i V_G. \tag{74}$$

With this maximum charge density, the surface potential collapses to the corresponding thermal-equilibrium value

$$\psi_s = 2\psi_B = \frac{2kT}{q}\ln\left(\frac{N_A}{n_i}\right) \tag{75}$$

which is to be avoided. Practical devices have maximum charge density of $\approx 10^{11}$ carriers/cm^2. A device 10 µm square thus can hold 10^5 carriers, and with a minimum detectable signal of $\approx$ 20 carriers, a dynamic range of $\approx 10^4$ can be achieved.

In addition to light, various sources for generation of dark current also supply charge to the surface and act as background noise. The total supply of charge density is given by the sum of dark current J_{da} and the photocurrent,

$$\begin{aligned}\frac{dQ_{\text{sig}}}{dt} &= J_{da} + J_{ph}\\ &= \frac{qn_iW_D}{2\tau} + \frac{qn_iS_o}{2} + \frac{qn_i^2L_D}{N_A\tau} + q\eta\Phi.\end{aligned} \tag{76}$$

Here the first three terms represent, in order, (1) generation in the depletion region, (2) generation at the surface, and (3) generation in the neutral bulk. The dark current also limits the maximum integration time to

$$t = \frac{Q_{max}}{J_{da}} \tag{77}$$

before it forces the system back to thermal equilibrium. Typical exposure time is in the range 0.1 to 100 ms. For detection of very weak signals, cooling is often required to minimize the dark current so that a longer integration time can be used. After the exposure period, the charge is transported to an amplifier by the CCD shift-register action. Such a mechanism is discussed in detail in the next section.

Because the CCDs can also be used as a shift register, there is great benefit to using this photodetector in an imaging-array system since the signals can be brought out sequentially to a single node, without complicated *x-y* addressing to each pixel. The detection mode of integrating charge over a long period of time enables detection of weaker signals. This is an important feature for astronomy imaging. In addition, the CCDs have the advantages of low dark current, low-noise, low-voltage operation, good linearity, and good dynamic range. The structure is simple and compact, stable and robust, and is compatible with MOS technology. These factors contribute to high yield, which enables the CCDs to be feasible in consumer products.

Different readout mechanisms for the line imager and the area imagers are shown in Fig. 24. A line imager with dual output registers has improved readout speed (Fig. 24a). Most-common area imagers use either interline-transfer (Fig. 24b) or frame-transfer (Fig. 24c) readout architecture. In the former, signals are transferred to the neighboring pixels, and they are subsequently passed along to the output register chain while the light-sensitive pixels start to collect charge for the next data. In the frame-transfer scheme, signals are shifted to a storage area away from the sensing area. The advantage of this compared to the interline transfer is a more efficient light-sensing area, but there is more image smear since CCDs continue to receive light as signal charges are passed through them. For both interline transfer and frame transfer, all columns advance their charge signals to the horizontal output register simultaneously, and the output register carries these signals out at a much higher clocking rate.

Charge-Injection Device. A charge-injection device (CID) does not necessarily have a different structure from a CCD. The difference lies in the readout mode. Instead of transferring the accumulated charge laterally, the charge-injection device releases the charge to the substrate by lowering the gate voltage. In an area-imaging system, *x-y* addressing of this photodetector is accomplished by implementing a two-well unit cell as shown in Fig. 25. With two closely spaced gates, the photogenerated charge can be shifted between the wells, controlled by the gate voltages. The charge is injected to the substrate only when both gate potentials are lowered and the semiconductor surface is driven into accumulation.

There are two readout mechanisms for the CIDs: sequential injection and parallel injection.[41] In the sequential-injection scheme, a pixel is selected when both gate

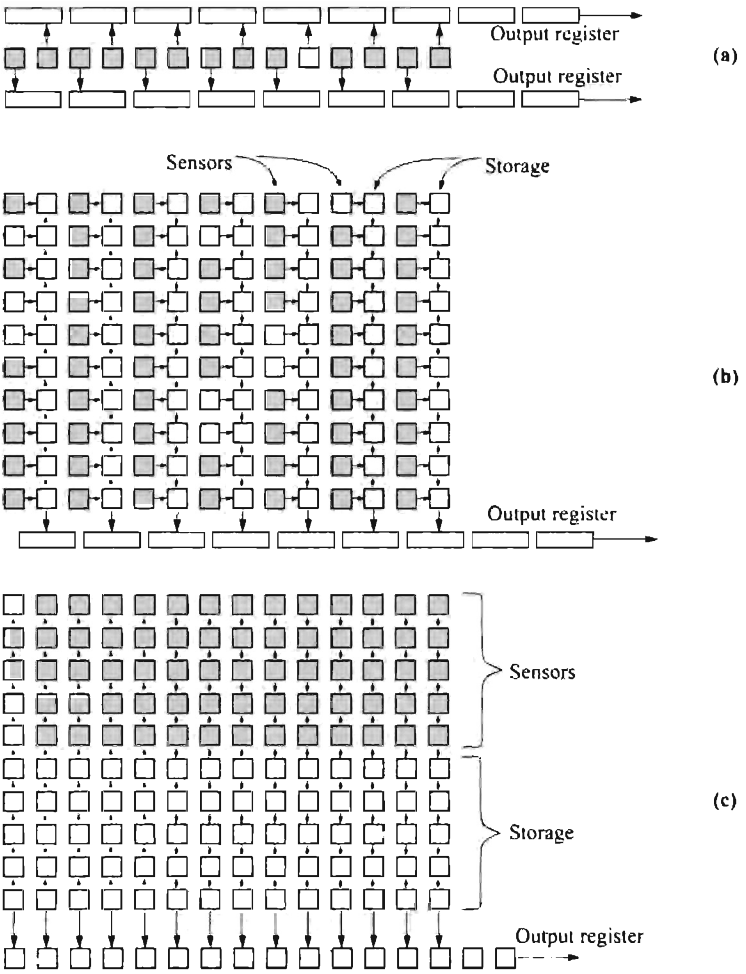

Output register
Output register
(a)
Sensors
Storage
(b)
Output register
Sensors
Storage
(c)
Output register

Fig. 24 Schematic layouts showing the readout mechanisms of (a) line imager with dual output registers, and area imagers with (b) interline transfer and (c) frame transfer. Gray pixels represent CCDs as photodetectors. The output register is usually clocked at a higher frequency than the internal transfer.

potentials are left to float, and as the charge is injected into the substrate, a displacement current can be sensed either at the substrate terminal or at the gate (Fig. 26a). In the parallel-injection scheme, an entire row is selected and all columns are read at the same time (Fig. 26b). A signal is detected when charge is transferred from one well (which has a higher gate voltage and/or thinner gate dielectric) to another within the unit cell. In such a readout as a displacement current in the gate, the charge is preserved.

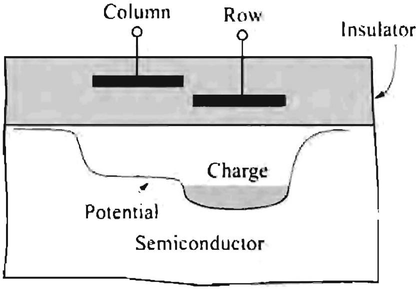


Fig. 25 Structure of charge-injection device with dual gates controlling two adjacent potential wells. Charge can be shifted between the wells or released to the substrate.

The CID area arrays have the advantage of random access capability. Transfer between cells is not necessary, and therefore transfer efficiency is not critical. The trade-off is higher power dissipation, which can be improved by an epitaxial substrate, larger noise due to large capacitance of the entire column, and the requirement of a better sensing amplifier due to weaker signals.

13.6.2 CCD Shift Register

In this section, we discuss the transfer of charge between CCDs. For optical imaging applications, the charge packets are formed as a result of electron-hole pair generation caused by light incident as previously discussed. For analog and memory devices, the charge packets are introduced by injection from a *p-n* junction at the vicinity of the CCD. Independent of the origin of the charge package, the transfer mechanisms are the same.

The CCD was invented by Boyle and Smith in 1970.[42] When CCDs are placed close together and with the proper sequence of gate voltages applied, minority-carrier

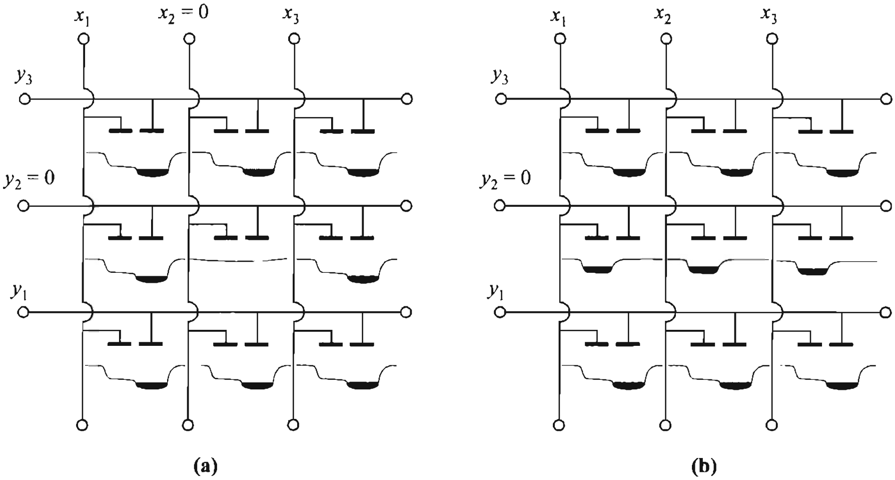


Fig. 26 Readout mechanisms in area charge-injection device arrays. (a) Sequential injection. (b) Parallel injection. In (a), $(x,y) = (2,2)$ is selected. In (b), the entire row of y_2 is selected.

charges at the surface can flow between devices and a simple shift register can be realized. Independently, the concept of an MOS bucket-brigade device (BBD) that performs similar functions was proposed by Sangster et al. around the same time.[43] CCD can be viewed as the integrated version of BBD. Most CCDs are made from the Si MOS system because of the good interfacial properties of thermally grown SiO_2, although in some specific applications, MIS structures, Schottky barriers, and heterojunctions on other semiconductors can also be used.

Figure 27 demonstrates the basic principle of charge transfer in a three-phase, n-channel CCD chain. The electrodes connected to the ϕ_1, ϕ_2, and ϕ_3 clock lines form the main body of the CCD. Figure 27b shows the clock waveforms and Fig. 27c illustrates the corresponding potential wells and charge distributions.

At $t = t_1$, clock line ϕ_1 is at a high voltage and ϕ_2 and ϕ_3 are at low voltages. The potential wells under ϕ_1 will be deeper than others. We assume that there is a signal charge at the first ϕ_1-electrode. At $t = t_2$, both ϕ_1 and ϕ_2 have high bias so charge start to transfer. At $t = t_3$, the voltage at ϕ_1 is returning to the low value while the ϕ_2-electrodes are still held at high voltage. The electrons stored under ϕ_1 are being emptied in this period. The remaining charge decay under the first node has a slowly falling edge, because the charge carriers require a finite time to transport across the width of the electrode. At $t = t_4$, the charge transfer is complete and the original charge packet

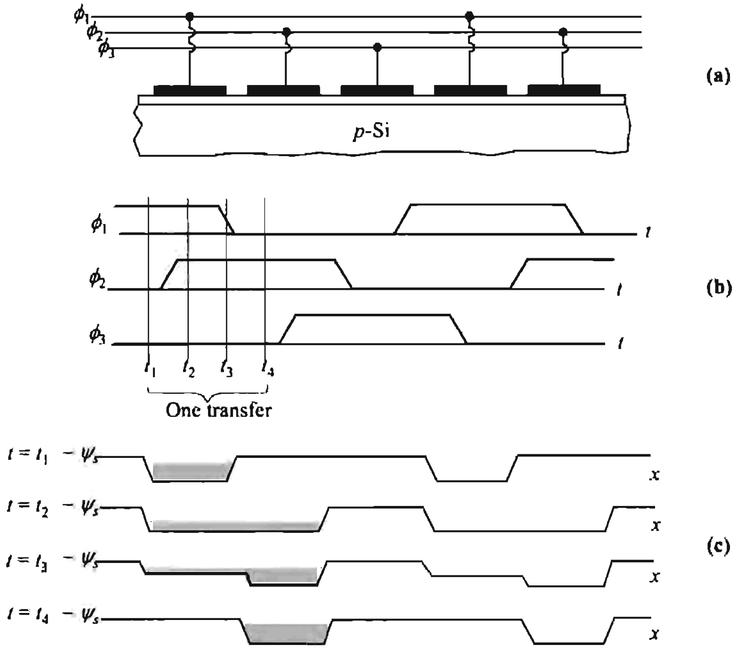


Fig. 27 Illustration of CCD charge transfer. (a) Application of 3-phase gate bias. (b) Clock waveforms. (c) Surface potential (and charge) vs. distance at different time.

is now stored under the first ϕ_2-electrode. This process will be repeated and the charge packet continues to shift to the right.

CCDs can be operated with two, three, or four phases, depending on the design of structures. Some representative structures are shown in Fig. 28. The spacing between CCDs should be small for efficient charge transfer. For two-phase operation, asymmetrical structures are required to define the direction of charge flow. Many electrode structures and clocking schemes have been proposed and implemented.[44]

Charge-Transfer Mechanisms. The three basic charge-transfer mechanisms are (1) thermal diffusion, (2) self-induced drift, and (3) fringing-field effect. For a small amount of signal charge, thermal diffusion is the dominant transfer mechanism. The total charge under the storage electrode decreases exponentially with time, and the time constant is given by[45]

$$\tau_{th} = \frac{4L^2}{\pi^2 D_n} \tag{78}$$

where L is the length of the electrode and D_n is the minority carrier diffusion constant.

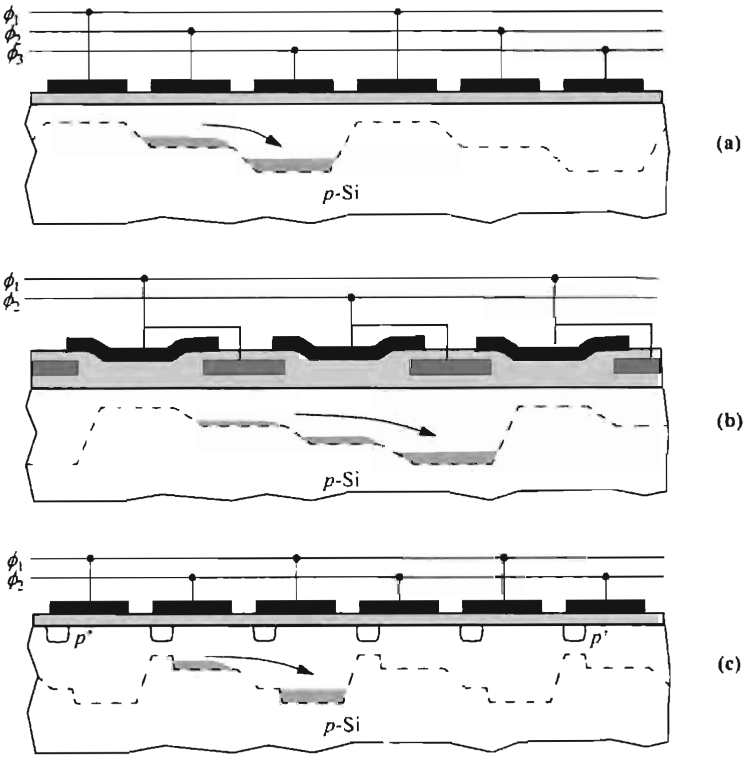


Fig. 28 CCD shift registers using (a) 3-phase single-level gate, (b) 2-phase with step oxides, (c) 2-phase with heavily doped pockets. Dashed lines indicate channel potential.

For a reasonably large charge packet, the transfer is dominated by the self-induced drift produced by electrostatic repulsion of the carriers. The magnitude of the self-induced longitudinal electrical field $\mathscr{E}_{xs}$ can be estimated by taking the gradient of the surface potential (which is assumed to vary linearly with the signal charge as given by Eq. 73)

$$\mathscr{E}_{xs} \approx \frac{1}{C_i} \frac{dQ_{\text{sig}}(x, t)}{dx} . \tag{79}$$

The decay of the initial charge packet due to the self-induced field is given by[46]

$$\frac{Q_{\text{sig}}(t)}{Q_{\text{sig}}(t=0)} = \frac{t_0}{t + t_0} \tag{80}$$

with

$$t_0 \equiv \frac{\pi L^2 C_i}{2\mu_n Q_{\text{sig}}} , \tag{81}$$

where μ_n is the mobility of the carriers.

The surface potential under the storage electrode is affected by the voltage applied to the adjacent electrodes due to two-dimensional coupling of the electrostatic potential. The applied voltage results in a surface electric field, even in the absence of signal charge at the interface. This fringing field is a function of the oxide thickness, electrode length, substrate doping, and gate voltage. It is a function of distance from the surface of the semiconductor, and it maximizes at a depth of $\approx L/2$. Because of this, BCCD can benefit much more from the fringing field than SCCD can. Figure 29 gives an example of this effect.[47] Because the fringing field is present even at very low charge concentration, the last bit of the signal charge will be transferred effectively by the fringing field.

We shall now define a transfer efficiency η, which is the ratio of charge transferred between electrodes:

$$\eta = 1 - \frac{Q_{\text{sig}}(t=T)}{Q_{\text{sig}}(t=0)} \tag{82}$$

where T is the total transfer period. A closely related concept is the transfer inefficiency ε, defined as

$$\varepsilon \equiv 1 - \eta = \frac{Q_{\text{sig}}(t=T)}{Q_{\text{sig}}(t=0)} . \tag{83}$$

Figure 29 shows that transfer efficiencies larger than 99.99% (or a transfer inefficiency less than 10^{-4}) can be obtained in the presence of the fringing field for clock frequencies of several tens of MHz. As frequencies become higher, gate length must be reduced to increase the fringing field.

The time-dependent surface potential and the transient behavior of the charge distribution have been computed using a two-dimensional model based on the charge continuity and current transport equation. Figure 30 shows a representative result.[48] Figure 30a shows that at the beginning of the charge-transfer process, the speed of the

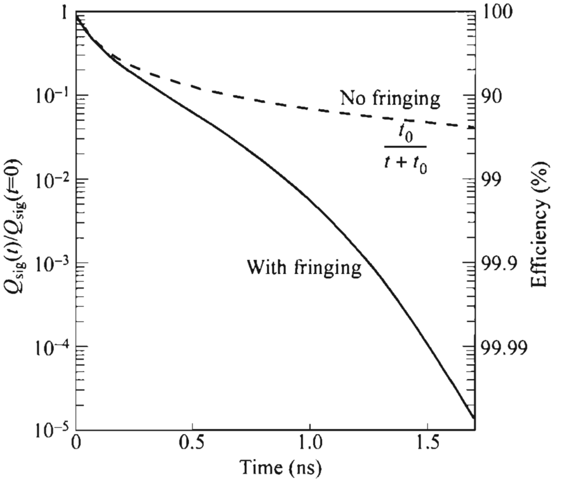


Fig. 29 Normalized remaining charge vs. time (4-μm gate length and 10^{15} cm^{-3} doping). The dashed line indicates how the charge transfer would proceed in the absence of a fringing field. (After Ref. 47.)

transfer is high because of a strong self-induced drift and a fringing field that results in a high drift velocity. After 0.8 ns, the surface potential changes very little, indicating that the amount of charge left to be transferred is small. The potential difference between the two neighboring potential wells after 0.8 ns (which is very close to the final potential difference when all the charges are completely transferred) is about 1.5 V. When the two potential wells approach each other, the transfer slows down considerably. Figure 30b shows the transient behavior of the charge distribution. The electrons under the storage gate-A are distributed more widely than the electrons under the transfer gate-B, because the fringing field near the edges of the potential wells forces electrons to move to the center of the wells. The fringing field under gate-B is stronger than that under gate-A. Therefore, electrons under gate-B are localized near the center of the gate. Also note from Fig. 30b that after 0.8 ns, about 99% of the electrons have been transferred.

In the discussion above, we considered only the free electrons in the conduction band. We have not considered the transition of charges in interface traps. Thus, the charge-transfer mechanism treated here is called the free-charge transfer model. The transfer efficiency at high frequencies can be described by this model and is limited by the clock rates for a given device. The maximum operating frequency can be well above 10 MHz for CCD with gate lengths smaller than 10 μm. At medium frequencies charge trapping at the interface traps determines the transfer efficiency.[49] When charge packets come in contact with empty interface traps, these traps are filled instantaneously, but when the signal charge has moved on, the interface traps release carriers with a whole spectrum of much slower time constants. Some trapped charges are released so rapidly from interface traps that they can move into the correct charge

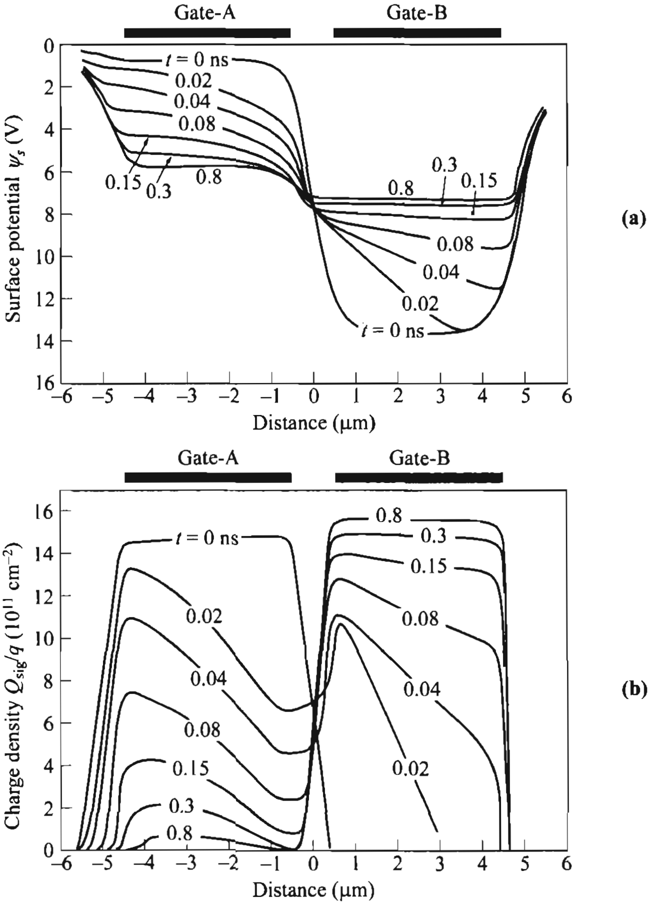


Fig. 30 (a) Time-dependent surface potential distribution under the storage and transfer gate (of length 4-μm). (b) Transient charge distribution under the gates. (After Ref. 48.)

packet, but others are released into trailing packets, which results in a charge loss from the leading charge packet and in a *tail* behind the last packet of a sequence. The transfer inefficiency due to interface traps is given by

$$\varepsilon \approx \frac{qkTD_{it}}{C_i \Delta \psi_s} \ln(N_p + 1) \tag{84}$$

where $\Delta\psi_s$ is the change in surface potential caused by the signal charge, D_{it} is the interface trap density, and N_p is the number of clock phases. To reduce ε, the interface trap density must be low. To avoid this effect, a background charge, called fat zero or a bias charge, can be used to fill these traps at all times, and this level of bias charge can be as large as 20%. The penalty is a reduced signal-to-noise ratio. Another way to get around the problem of interface traps is to use buried-channel CCD.

There are many other factors contributing to the transfer inefficiency. Among them are the nature of the exponential decay of charge during transfer by diffusion and fringing-field drift, and the finite transfer time within the clock period. Efficient transfer can also be hindered by barrier hump in the gap between devices.

Frequency Limitations. The choice of the period (or frequency) of the clock signal is limited by three factors. First, it has to be long enough for reasonably complete transfer of charge. Second, it has to be much shorter than the thermal-relaxation time to minimize minority carriers generated from dark current. Especially for analog signals, the clock period has to be small enough to avoid signal loss. Third, the clock period has to be small compared to the period of the analog signal ($1/f$) to be transmitted. These place both lower and upper limits on the frequency of operation.

At low clock frequencies, the frequency limit is determined by the dark current. The dark-current density J_{da} can be expressed as[45]

$$J_{da} = \frac{qn_iW_D}{2\tau} + \frac{qS_0n_i}{2} + \frac{qD_n}{L_n}\frac{n_i^2}{N_A} , \tag{85}$$

where the first term on the right-hand side is the bulk generation inside the depletion region, the second term is the surface generation current, and the last term is the diffusion current at the edge of the depletion region (τ is the minority-carrier lifetime and S_0 the surface generation/recombination velocity).

The low-frequency limit of a CCD can be estimated by comparing the charge accumulated from dark current with the signal charge. If a CCD is clocked continuously at a constant frequency f, the output signal due to dark current is[45]

$$Q_{da} = \frac{J_{da}N}{N_pf} \tag{86}$$

where N is the number of electrodes, and N_p the number of phases. The maximum signal charge a CCD can handle is

$$Q_{\max} = C_i\Delta\psi_s \tag{87}$$

where $\Delta\psi_s$ is the maximum surface potential change due to the maximum signal charge. Thus the ratio of noise background to signal is

$$\frac{Q_{da}}{Q_{\max}} = \frac{J_{da}N}{N_pfC_i\Delta\psi_s} . \tag{88}$$

The low-frequency degradation of frequency response is due to the buildup of dark current in the charge packets, which distort the size of the signal charge. To improve the low-frequency response, one must reduce all the dark-current components in Eq. 85 by having a long minority-carrier lifetime, large diffusion length, and low surface recombination velocity.

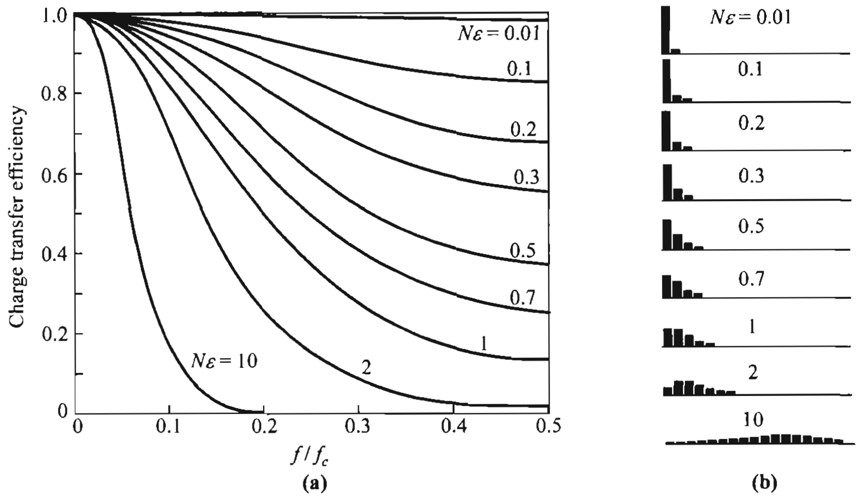
At high frequencies, transfer efficiency falls rapidly because there is not enough time to allow for complete charge transfer. To extend high-frequency performance, one can reduce the gate length (L), maximize surface mobility (by using electrons instead of holes in the charge packet), and minimize electrode spacings. The higher electron mobility in GaAs makes it possible to design ultra-high-speed CCDs. A heterojunction GaAs CCD has been operated up to an 18 GHz clock frequency.[50] The frequency dependence of the output efficiency, normalized to ε at the clock frequency f_c, is given by[51]

$$\frac{Q_{\text{sig}}(\text{output})}{Q_{\text{sig}}(\text{input})} = \exp\left[-N\varepsilon\left\{1 - \cos\left(\frac{2\pi f}{f_c}\right)\right\}\right], \qquad f < f_c. \tag{89}$$

Equation 89 is plotted in Fig. 31a.

Transfer inefficiency can introduce an extra phase delay. Figure 31b shows degradation of a single charge packet as a function of the $N\varepsilon$ product.[44] One can see the spreading of an individual charge packet into the trailing charge packets for larger values of $N\varepsilon$. The left-most cell in each frame represents the position where the original charge packet is expected to appear in an ideal CCD. Charge delayed by transfer inefficiency emerges in later time slots, shown toward the right. For $N\varepsilon \geq 1$ the inadequacy of the transfer efficiency is clear, because the main amount of charge no longer appears in the leading station.

Buried-Channel CCD. In the surface-channel CCD (SCCD), minority-carrier charge packets are moved along the surface of the semiconductor. One major limitation of this CCD is the effect of interface traps. To circumvent this problem and improve transfer efficiency, the buried-channel CCD (BCCD) was proposed in which the charge packets do not flow at the semiconductor surface; instead, they are confined to a channel that lies beneath the surface.[52] The BCCD has the potential of eliminating the interface trapping. A schematic cross-sectional view of a BCCD is shown in Fig. 23b.[53] It consists of an opposite-type (n-type) semiconductor layer on a p-type substrate. When no signal charge is present, the narrow n-type region is fully depleted



(a) (b)

Fig. 31 (a) Effect of transfer inefficiency product $N\varepsilon$ on frequency response. (b) Degradation of signals in successive cells from a single charge packet. (After Ref. 44.)

under a positive voltage pulse applied to the gate electrode. As signal charges are introduced, they will be stored in the buried channel. Because the signal charge is away from the surface, it has the advantages of higher mobility, less charge loss due to interface traps, and higher fringing fields for charge transfer. The penalty is less charge-handling capability because the charge is farther away from the gate, and, thus, less coupling.

Figure 32 shows a two-dimensional calculation of potential along the buried channel. For comparison, it also shows the potential plot for a surface device. Clearly, the BCCD has a greater potential gradient under the transferring electrode, which helps to speed up charge transfer. A transfer inefficiency to 10^{-4} to 10^{-5} is readily achieved in the BCCD, and is an order of magnitude smaller than a typical SCCD having the same device geometry.

13.6.3 CMOS Image Sensor

For consumer imaging products such as digital cameras and video recorders, the CCD image sensor has been dominating the market. However, this huge market has been taken away increasingly by CMOS image sensor since the late 1990s.[54] Even though in a CMOS image sensor there is little that is new in the photodetector, it is worthy of mentioning here because of the fast growth in using this option to replace CCD. The novelty lies in the increasing integration of more functionality within each pixel, taking advantages of the conventional CMOS scaling and inexpensive technology. Conversely, CCD requires different process optimization, so CCD systems which include CMOS circuitry are naturally more expensive.

The CMOS image sensor is not really a photodetector alone, but is an architecture of imaging followed by some functionality carried out within the pixel. The main three schemes of CMOS image sensors are shown in Fig. 33. These are called PPS (passive pixel sensor), APS (active pixel sensor), and DPS (digital pixel sensor). Each

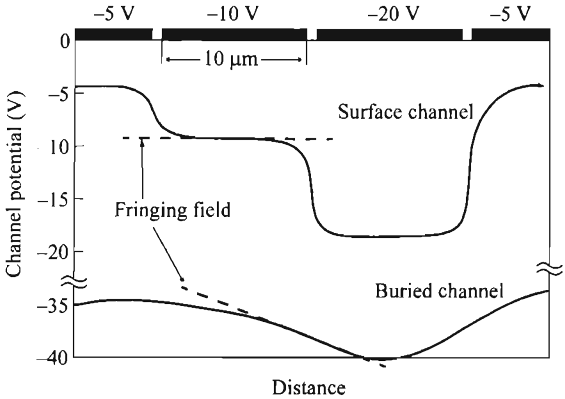


Fig. 32 Two-dimensional calculation of the potential along a BCCD. Higher fringing field (slope) for BCCD compared to SCCD is shown. (After Ref. 53.)

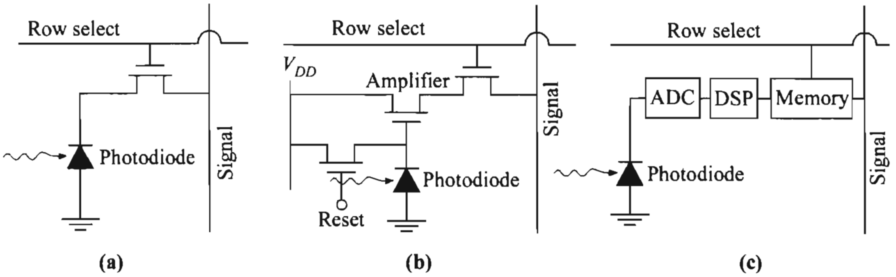


Fig. 33 Versions of CMOS image sensors: (a) PPS (passive pixel sensor), (b) APS (active pixel sensor), and (c) DPS (digital pixel sensor).

of these contains a photodetector which is commonly a *p-n* junction photodiode. But other options considered are p^+-*n*-*p* pinned diode where the middle layer is fully depleted like a planar-doped barrier diode,[55] and photogate which is similar to a CCD. These configurations have increasing area for each pixel, but more and more functions are added to the pixel level.

The PPS is the most-basic form of imaging array where in each pixel a select transistor controls each photodetector. The advantage is many cells in a row are accessed at the same time, as in a memory array, so the speed is higher than CCD whose read out is serial in nature. Its larger size is the penalty. The APS is the most-common configuration at the present time, where in each pixel, in addition to the photodiode and the select transistor, there are an amplifier whose gate is fed the photocurrent, and a reset transistor. Finally in the DPS, there is an analog-to-digital converter (ADC) after which digital signal processing (DSP), such as automatic gain control, can be performed within each pixel. Note that in both APS and DPS, signal charge is not lost during sensing as in CCD and PPS.

Compared to CCD, the advantages of the CMOS image sensor include higher speed due to random-access capability, larger signal-to-noise ratio, lower power due to low voltage requirement, and low cost because of main-stream technology. The CCD maintains some advantages such as small pixel size, low-light sensitivity, and high dynamic range.

13.7 METAL-SEMICONDUCTOR-METAL PHOTODETECTOR

The metal-semiconductor-metal (MSM) photodetector was proposed and demonstrated by Sugeta et al. in 1979.[56–57] The structure of the MSM photodetector, as shown in Fig. 34, is basically two Schottky barriers connected back-to-back, on a coplanar surface. The concept of adding a thin barrier-enhancement layer to reduce the dark current has been proven to be beneficial since it was introduced in 1988,[58–59] and most of the recent structures have incorporated this layer. The metal contacts usually have the shape of interdigitated stripes. Light is received at the gap between

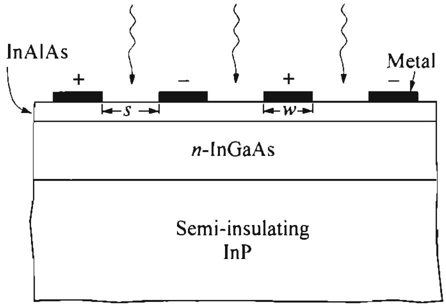


Fig. 34 The MSM photodetector is consisted of planar interdigitated metal-semiconductor contacts. The top layer (InAlAs) servers to reduce the dark current by providing a higher barrier height.

the metal contacts, and the MSM photodetector avoids absorption of light by the metal layer as in a conventional Schottky-barrier photodiode. For more complete light absorption, the active layer has a thickness slightly larger than the absorption length ($1/\alpha \approx 1$ μm) and has a low doping of $\approx 1\times10^{15}$ cm^{-3} for low capacitance. InGaAs received the most attention for applications in the 1.3–1.5 μm range, which has optimum performance for optical fibers.

In typical operations, the photocurrent first rises with voltage and then becomes saturated. The increase of photocurrent at low bias is due to the expansion of the depletion region in the reverse-biased Schottky junction, and the internal quantum efficiency is improved. The voltage at which the photocurrent saturates corresponds to the flat-band condition in which the electric field at the anode becomes zero (Fig. 35).[60] At this point, the quantum efficiency can be close to 100%. This condition can be estimated by a one-dimensional depletion equation

$$V_{FB} \approx \left(\frac{qN}{2\varepsilon_s}\right)s^2 \tag{90}$$

where N is the doping and s is the spacing between fingers. (Equation 90 is at punch-through when the depletion widths consume the entire spacing s, and it occurs before flat-band.) Operation beyond punch-through also has the advantage of minimum capacitance. Note that in the MSM photodetector, generation is via band-to-band excitation, but not using the option of photoexcitation over the barrier as in a regular metal-semiconductor photodiode (Fig. 11b).

Internal photocurrent gain is sometimes observed in the MSM photodetector. One explanation of the gain is photoconductivity, caused by long-lifetime traps located either within the barrier-enhancement layer or at the heterointerface. Another theory is that when photogenerated holes are accumulated at the valence-band peak near the cathode, these positive charges increase the field across the wide-energy-gap barrier-enhancement layer and induce a larger electron tunneling current. A similar effect can be true for electrons accumulated near the anode, and the hole tunneling current is enhanced. This mechanism is somewhat similar to a phototransistor. In any case, there has been an effort to eliminate this gain because the gain mechanism slows down the response time of the photodetector, especially the turn-off process.

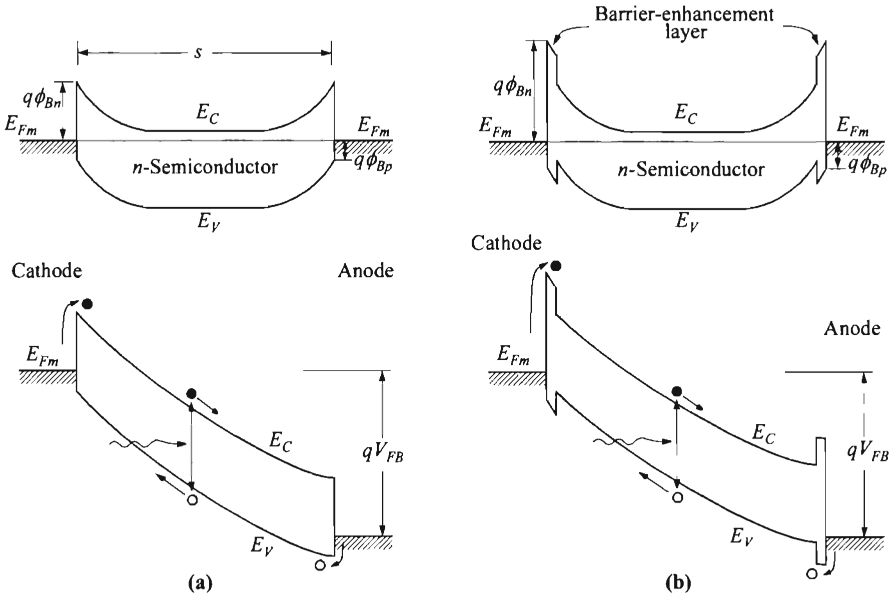


Fig. 35 Energy-band diagrams of MSM photodetectors at equilibrium and under bias at flat-band. (a) Without and (b) with the barrier-enhancement layer.

The main drawback of the MSM photodetector is high dark current, due to the Schottky-barrier junction. This is especially serious for low-bandgap material where long-wavelength detection is needed. However, a barrier-enhancement layer can drastically reduce the dark current of a narrow-energy-gap semiconductor such as InGaAs. By inserting this layer of wider energy gap, the barrier height becomes much larger. This layer ranges in thickness from 30 to 100 nm. The barrier-enhancement layer can be graded in composition to avoid carrier trapping at the band-edge discontinuity (see Fig. 35b near the cathode).

Since the MSM photodetector has two Schottky barriers connected back-to-back, bias of any polarity will put one Schottky barrier in the reverse direction (cathode) and the other in the forward direction (anode). Energy-band diagrams between the two metal contacts through the active layer are shown in Fig. 35. The most-common dark I-V characteristics have current saturation at a low voltage and is typical of thermionic-emission current. Considering both electron and hole current components here, the saturation current has the general expression[60]

$$I_{da} = A_1 A_n^* T^2 \exp\left(\frac{-q\phi_{Bn}}{kT}\right) + A_2 A_p^* T^2 \exp\left(\frac{-q\phi_{Bp}}{kT}\right), \tag{91}$$

where A_1 and A_2 are the anode and cathode contact areas, A_n^* and A_p^* are the effective Richardson constants for electrons and holes respectively. For higher bias, the current can continue to rise with bias. This nonsaturating current can be due to image-force lowering that modifies the barrier height, or from tunneling through the barrier.

The MSM photodetector has the primary advantages of high speed and compatibility with FET technology. Its simple, planar structure is easy to integrate with FETs in a single chip. The MSM photodetector has very low capacitance per area, because of two-dimensional effects on a semi-insulating substrate. This is especially advantageous for detectors requiring large light-sensitive areas. Compared to a *p-i-n* photodiode or a Schottky-barrier photodiode of similar quantum efficiency, its capacitance is reduced to about half. With such a small capacitance, the *RC* charging time and speed are much improved. Speed is also determined by the transit time, which is directly proportional to the spacing dimension. For this reason, a small spacing is preferable for speed consideration. A bandwidth of higher than 100 GHz has been reported.[61]

To understand the speed optimization, an example showing the theoretical analysis of an MSM photodetector is given in Fig. 36. In this particular example, the speed is not very high due to the materials and structure chosen. Nevertheless it gives some insight into the factors affecting the speed performance. The speed can be limited by *RC* time constant and the transit time. The bandwidth due to *RC* time constant is given by[62]

$$f_{RC} = \frac{1}{2\pi(R_L + R_s)C} \tag{92}$$

where R_L is the load resistance (= 50 Ω) and R_s the series resistance. The capacitance is given by

$$C = \frac{K(\kappa)}{K(\kappa')} \frac{\varepsilon_0 A(1 + K_s)}{(s + w)} \tag{93}$$

where A is the area of the contacts, K_s the relative dielectric constant of the semiconductor, and $K(\kappa)$ the complete elliptic integral of the first kind

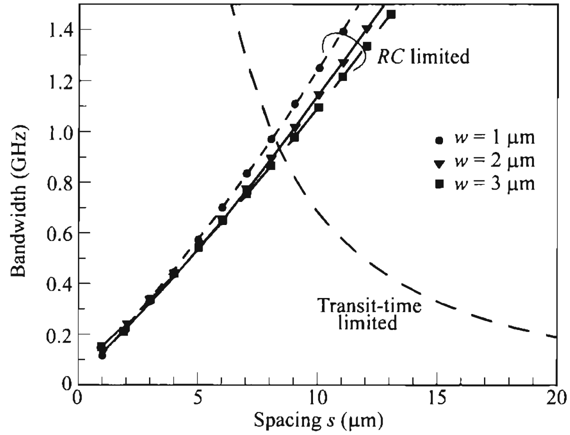


Fig. 36 Theoretical bandwidth of MSM photodetector for various finger width w and spacing s. Example assumes 1-μm of $In_{0.53}Ga_{0.47}As$ active layer. (After Ref. 62.)

$$K(\kappa) = \int_0^{\pi/2} \frac{1}{\sqrt{1-\kappa^2 \sin^2 \varphi}} d\varphi, \tag{94}$$

with

$$\kappa = \tan^2\left[\frac{\pi w}{4(s+w)}\right] \quad , \quad \kappa' = \sqrt{1-\kappa^2}. \tag{95}$$

The transit-time-limited bandwidth is given by

$$f_{tr} = \frac{0.44}{\sqrt{2}}\left(\frac{v_s}{s}\right) \tag{96}$$

where it is assumed carriers travel with the saturation velocity v_s. It is seen in Fig. 36 that the speed is not sensitive to the finger width w. For the spacing dimension, the *RC* time constant and the transit time have opposite trend, and for this example the optimized spacing is ≈ 8 μm.

13.8 QUANTUM-WELL INFRARED PHOTODETECTOR

Infrared absorption within the conduction band or the valence band, instead of band-to-band, in a quantum well was first studied during 1983–1985.[63–65] The first functional quantum-well infrared photodetector (QWIP), based on bound-to-bound intersubband transition in a GaAs/AlGaAs heterostructure, was realized by Levine et al.[66] and Choi et al.[67] in 1987. The same group also presented improved detector results on bound-to-continuum transition in 1988.[68] Another type of transition, bound-to-miniband had been observed in 1991.[69]

The structure of a QWIP using a GaAs/AlGaAs heterostructure is shown in Fig. 37. The quantum-well layers, in this case GaAs, have a thickness of about 5 nm

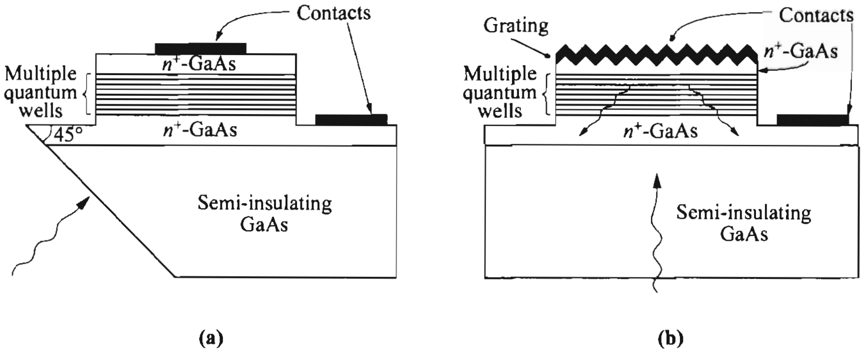


Fig. 37 Structures of GaAs/AlGaAs QWIPs showing approaches to couple light to the heterointerface at a critical angle. (a) Light is incident normal to a polished facet making a 45° angle to the quantum well. (b) A grating is used to refract light coming from the substrate.

and are usually doped to *n*-type in the 10^{17} cm^{-3} range. The barrier layers are undoped and have a thickness in the range of 30–50 nm. A typical number of periods is between 20 and 50.

For quantum wells formed by direct-bandgap materials, incident light normal to the surface has zero absorption because intersubband transitions require that the electric field of the electromagnetic wave has components normal to the quantum-well plane. This polarization selection rule demands other approaches to couple light to the light-sensitive area, and two popular schemes are shown in Fig. 37. In Fig. 37a, a polished 45°-facet is made at the edge adjacent to the detector. Notice that the wavelength of interest is transparent to the substrate. In Fig. 37b, a grating on the top surface refracts light back to the detector. Alternatively, a grating can be made on the substrate surface to scatter the incoming light. This selection rule, however, does not apply to *p*-type quantum wells or wells formed by indirect-bandgap materials such as SiGe/Si and AlAs/AlGaAs heterostructures.

The QWIP is based on photoconductivity due to intersubband excitation. The three types of transitions are depicted in Fig. 38. In the bound-to-bound transition, both quantized energy states are confined and below the barrier energy. A photon excites an electron from the ground state to the first bound state and the electron subsequently tunnels out of the well. In the bound-to-continuum (or bound-to-extended) excitation, the first state above the ground state is over the barrier and excited electrons can escape the well more easily. This bound-to-continuum excitation is more promising in that it has higher absorption, broader wavelength response, lower dark current, higher detectivity, and requires lower voltage. In the bound-to-miniband transition, a miniband is present because of the superlattice structure. QWIPs based on this have shown great promise for focal-plane array imaging sensor system applications.

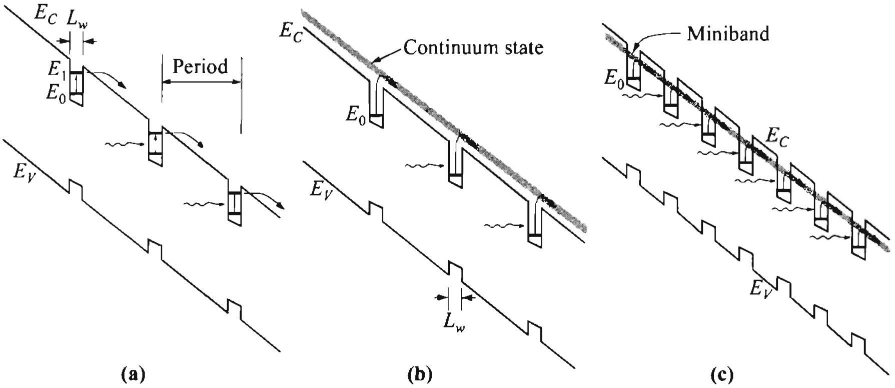


Fig. 38 Energy-band diagrams of QWIPs under bias showing (a) bound-to-bound intersubband transition, (b) bound-to-continuum transition, and (c) bound-to-miniband transition in superlattice.

The general QWIP *I-V* characteristics are similar to that of a regular photoconductor. Asymmetric characteristics might occur due to band bending arising from a dopant-migration effect in the quantum wells. The photocurrent is given by the same general expression used for photoconductors,

$$I_{ph} = q\Phi_{ph}\eta G_a \tag{97}$$

where Φ_{ph} is the total photon flux (s^{-1}) and G_a is the optical gain. The quantum efficiency η is different from a photoconductor since in the QWIP, light absorption and carrier generation occur only in the quantum wells but not homogeneously throughout the structure. It is given by

$$\eta = (1-R)[1-\exp(-N_{op}\alpha N_w L_w)]E_p P \tag{98}$$

where R is the reflection, N_{op} is the number of optical passes, and N_w is the number of quantum wells, each of length L_w. The escape probability E_p is a function of bias which measures the extraction of the excited carriers out of the quantum wells.[70] The polarization correction factor P for GaAs is 0.5 for *n*-type quantum wells, and 1.0 for *p*-type quantum wells. The absorption coefficient α is a function of the incident angle and is proportional to $\sin^2\theta$, where θ is the angle between light propagation and the normal of the quantum-well plane.

The photoconductive gain has been derived to be[71–72]

$$G_a = \frac{1}{N_w C_p} \tag{99}$$

where C_p is the capture probability of an electron traversing a quantum well, given by

$$C_p = \frac{t_p}{\tau} = \frac{t_t}{N_w \tau}. \tag{100}$$

t_p is the transit time across a single period of the structure and t_t is the transit time across the entire QWIP active length L (wells and barriers). Combining Eqs. 99 and 100 yields

$$G_a = \frac{\tau}{t_t} \tag{101}$$

and it is similar to the gain of a standard photoconductor. For carriers in the mobility regime (before velocity saturation),

$$t_t = \frac{L}{\upsilon_d} = \frac{L^2}{\mu V} \tag{102}$$

where a uniform field is assumed across the entire length L, giving

$$G_a = \frac{\tau\mu V}{L^2}. \tag{103}$$

The dark current of a QWIP is due to thermionic emission over the quantum-well barriers and thermionic-field emission (thermally assisted tunneling) near the barrier peaks. Since this photodetector aims at wavelengths from ≈ 3 μm to about 20 μm, the barriers forming the wells have to be small, around 0.2 eV. In order to limit the dark current, the QWIP has to be operated at low temperatures, in the range 4–77 K.

The QWIP is an attractive alternative for long-wavelength photodetectors that use HgCdTe material, which has problems of excessive tunneling of dark current and reproducibility in precise composition to give the exact energy gap. It is compatible with GaAs technology and circuits for monolithic integration. The detection wavelength range is also tunable by the quantum-well thickness. Long-wavelength capability close to 20 μm has been demonstrated.[70] The QWIP can be applied in focal-plane arrays for two-dimensional imaging. Examples are thermal and terrestrial imaging. The QWIP is also known for its high-speed capability and fast response. This is due to its intrinsic short carrier lifetime in the quantum wells which is in the order of 5 ps. One difficulty with the QWIP, at least for *n*-type GaAs wells, is detecting normal-incidence light. This makes coupling of light to the photodetector difficult.

13.9 SOLAR CELL

13.9.1 Introduction

Solar cells, at the present time, furnish the most-important long-duration power supply in small-scale terrestrial and space applications such as satellites and space vehicles. As worldwide energy demand increases, conventional energy resources such as fossil fuels, will be exhausted within the next century. Therefore, we must develop and use alternative energy resources, especially our only long-term natural resource—the sun. The solar cell is considered a major candidate for obtaining energy from the sun, since it can convert sunlight directly to electricity with high conversion efficiency (as opposed to extracting thermal energy). It can provide nearly permanent power at low operating cost, and is virtually free of pollution. Recently, research and development of low-cost flat-panel solar panels, thin-film devices, concentrator systems, and many innovative concepts have increased. In the near future, the costs of small solar-power modular units and solar-power plants will be economically feasible for large-scale production and use of solar energy.

The photovoltaic effect, the generation of voltage when a device is exposed to light, was discovered by Becquerel in 1839, in a junction formed between an electrode and an electrolyte.[73] Since then there had been reports of similar effects on different solid-state devices. The first photovoltaic effect of substantial EMF voltage was observed by Ohl on a silicon *p-n* junction in 1940.[74–75] The photovoltaic effect on Ge was reported by Benzer in 1946[76] and by Pantchechnikoff in 1952.[77] It was not until 1954 that the solar cell received much increased interest, initiated by the works of Chapin et al. on single-crystal silicon cells[78] and of Reynolds et al. on cadmium sulfide cells.[79] To date, solar cells have been made in many other semiconductors, using various device configurations, and employing single-crystal, polycrystal, and amorphous thin-film structures.

A solar cell is similar to a photodiode. The photodiode can be operated in a photovoltaic mode, that is, it is unbiased and connected to a load impedance similar to a solar cell. However, the device designs are fundamentally different. For a photodiode only a narrow wavelength range centered at the optical signal wavelength is important, whereas for a solar cell, wide spectral response over a broad solar wavelength range is required. Photodiodes are small to minimize junction capacitance, while solar cells are large-area devices. One of the most-important figures of merit for photodiodes is the quantum efficiency, whereas the main concern for solar cells is the power conversion efficiency (power delivered to the load per incident solar energy).

13.9.2 Solar Radiation and Ideal Conversion Efficiency

Solar Radiation. The radiative energy output from the sun derives from a nuclear fusion reaction. In every second about 6×10^{11} kg of H_2 is converted to He, with a net mass loss of about 4×10^3 kg, which is converted through the Einstein relation ($E = mc^2$) to 4×10^{20} J. This energy is emitted primarily as electromagnetic radiation in the ultraviolet to infrared and radio spectral ranges (0.2 to 3 µm). The total mass of the sun is now about 2×10^{30} kg, and a reasonably stable life with a nearly constant radiative energy output over 10 billion (10^{10}) years is projected.

The intensity of solar radiation in free space at the average distance of the earth from the sun has a value of 1,353 W/m^2. The atmosphere attenuates the sunlight when it reaches the earth's surface, mainly due to water-vapor absorption in the infrared, ozone absorption in the ultraviolet, and scattering by airborne dust and aerosols. The degree to which the atmosphere affects the sunlight received at the earth's surface is quantified by the *air mass*. The secant of the angle between the sun and the zenith (sec θ) is defined as the air mass (AM) number and it measures the atmospheric path length relative to the minimum path length when the sun is directly overhead. The AM0 thus represents the solar spectrum outside the earth's atmosphere. The AM1 spectrum represents the sunlight at the earth's surface when the sun is at zenith, and the incident power is about 925 W/m^2. The AM2 spectrum is for $\theta = 60°$ and has an incident power of about 691 W/m^2, and so on.

Figure 39 shows the solar spectrums at various AM conditions. The upper curve is the AM0 condition which can be approximated by a 5,800 K black-body radiation, as shown by the dashed curve. The AM0 spectrum is the relevant one for satellite and space-vehicle applications. The AM1.5 conditions (with sun at 45° above the horizon) represent a satisfactory energy-weighted average for terrestrial applications. For solar-cell energy conversion, each photon produces an electron-hole pair, so the solar power has to be converted to photon flux. The photon flux density per unit energy for AM1.5 is shown in Fig. 40 together with that for AM0. To convert the wavelength to photon energy, we use the relationship

$$\lambda = \frac{c}{\nu} = \frac{1.24}{h\nu\,(\text{eV})} \quad \mu\text{m}. \tag{104}$$

The total incident power for AM1.5 is 844 W/m^2.

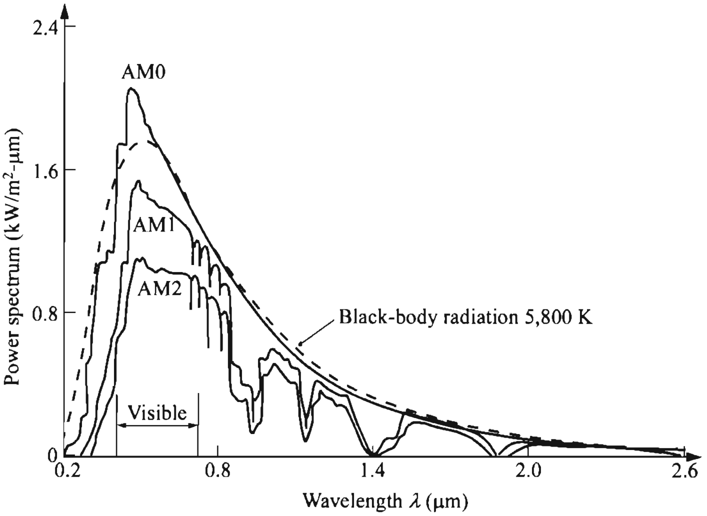


Fig. 39 Solar spectrum at different air-mass conditions. (After Ref. 80.)

Ideal Conversion Efficiency. The conventional solar cell, typically a *p-n* junction, has a single bandgap E_g. When the cell is exposed to the solar spectrum, a photon with energy less than E_g makes no contribution to the cell output (neglecting phonon-assisted absorption). A photon with energy greater than E_g contributes an electric charge to the cell output, and the excess energy over E_g is wasted as heat. To derive the ideal conversion efficiency, we shall consider the energy band of the semicon-

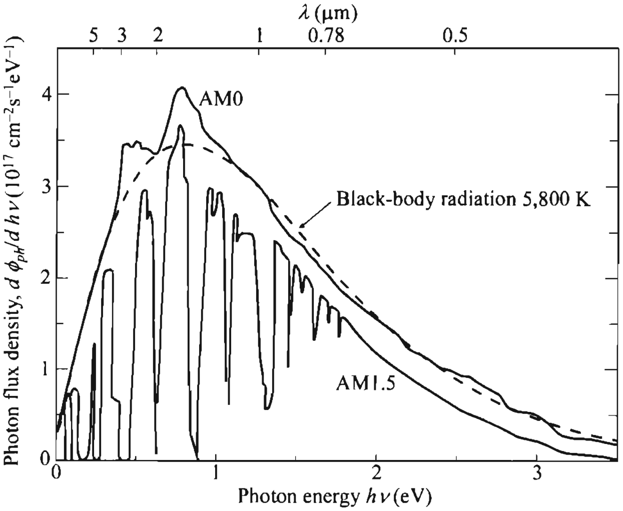


Fig. 40 Solar spectrum in photon flux density per photon energy for AM0 and AM1.5 conditions. (After Ref. 81.)

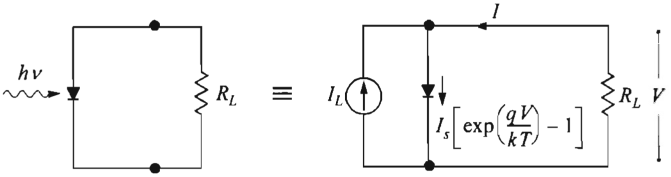


Fig. 41 Idealized equivalent circuit of solar cell under illumination.

ductor used. The solar cell is assumed to have ideal diode I-V characteristics. The equivalent circuit is shown in Fig. 41 where a constant-current source of photocurrent is in parallel with the junction. The source I_L results from the excitation of excess carriers by solar radiation; I_s is the diode saturation current as derived in Chapter 2, and R_L is the load resistance.

To obtain the photocurrent I_L, we need to integrate the total area under the graph shown in Fig. 40, that is,

$$I_L(E_g) = Aq\int_{h\nu=E_g}^{\infty} \frac{d\phi_{ph}}{dh\nu}d(h\nu). \tag{105}$$

The result is shown in Fig. 42 as a function of the bandgap of the semiconductor. For the photocurrent consideration, the smaller bandgap the better because more photons are collected.

The total I-V characteristics of such a device under illumination is simply a summation of the dark current and the photocurrent, given as

$$I = I_s\left[\exp\left(\frac{qV}{kT}\right) - 1\right] - I_L. \tag{106}$$

From Eq. 106 we obtain the open-circuit voltage by setting $I = 0$:

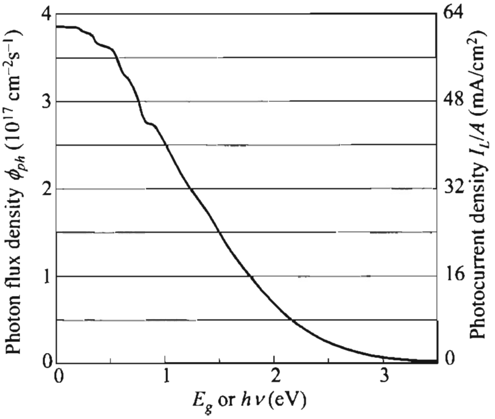


Fig. 42 Total number of photons in the solar spectrum (of AM1.5) above an energy value, contributing to the maximum photocurrent for a solar cell made with a specific E_g. (After Ref. 81.)

$$V_{oc} = \frac{kT}{q}\ln\left(\frac{I_L}{I_s}+1\right) \approx \frac{kT}{q}\ln\left(\frac{I_L}{I_s}\right). \tag{107}$$

Hence for a given I_L, the open-circuit voltage increases logarithmically with decreasing saturation current I_s. For a regular *p-n* junction, the ideal saturation current is given by

$$I_s = AqN_CN_V\left(\frac{1}{N_A}\sqrt{\frac{D_n}{\tau_n}}+\frac{1}{N_D}\sqrt{\frac{D_p}{\tau_p}}\right)\exp\left(\frac{-E_g}{kT}\right). \tag{108}$$

As seen, I_s decreases exponentially with E_g. So to obtain a large V_{oc}, a large E_g is required. Qualitatively, we know the maximum V_{oc} is the built-in potential of the junction, and the maximum built-in potential is close to the energy gap.

A plot of Eq. 106 is given in Fig. 43. The curve passes through the fourth quadrant and, therefore, power can be extracted from the device to a load. By properly choosing a load, close to 80% of the product $I_{sc}V_{oc}$ can be extracted. Here I_{sc} is the short-circuit current which is equal to the photocurrent derived. The shaded area is the maximum power output. We also define in Fig. 43 the quantities I_m and V_m that correspond to the current and voltage, for the maximum power output P_m ($= I_mV_m$).

To derive the maximum-power operating point, the output power is given by

$$P = IV = I_sV\left[\exp\left(\frac{qV}{kT}\right)-1\right]-I_LV. \tag{109}$$

The condition for maximum power can be obtained when $dP/dV = 0$, or

$$I_m = I_s\beta V_m\exp(\beta V_m) \approx I_L\left(1-\frac{1}{\beta V_m}\right), \tag{110}$$

$$V_m = \frac{1}{\beta}\ln\left[\frac{(I_L/I_s)+1}{1+\beta V_m}\right] \approx V_{oc}-\frac{1}{\beta}\ln(1+\beta V_m), \tag{111}$$

where $\beta \equiv q/kT$. The maximum power output P_m is then

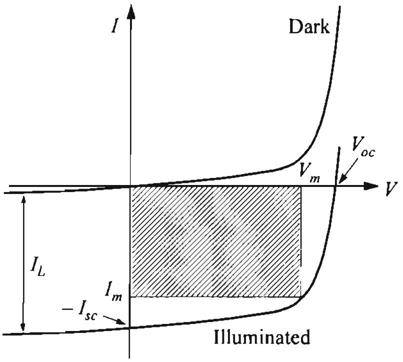


Fig. 43 *I-V* characteristics of solar cell under illumination. Determination of maximum power output is indicated.

$$P_m = I_m V_m = F_F I_{sc} V_{oc} \approx I_L\left[V_{oc} - \frac{1}{\beta}\ln(1+\beta V_m) - \frac{1}{\beta}\right], \qquad (112)$$

where the fill factor F_F measures the sharpness of the curve and is defined as

$$F_F \equiv \frac{I_m V_m}{I_{sc} V_{oc}}. \qquad (113)$$

In practice, a good fill factor is around 0.8. The ideal conversion efficiency is the ratio of the maximum power output to the incident power P_{in},

$$\eta = \frac{P_m}{P_{in}} = \frac{I_m V_m}{P_{in}} = \frac{V_m^2 I_s (q/kT)\exp(qV_m/kT)}{P_{in}}. \qquad (114)$$

Theoretically, the ideal efficiency can be calculated. We have shown that the photocurrent increases with smaller E_g. On the other hand, the voltage increases with E_g by having a small saturation current. So to maximize the power, there exists an optimum value for the bandgap E_g. Furthermore, by using the ideal saturation current of Eq. 108 in relation to E_g, the theoretical maximum conversion efficiency can be calculated. The ideal efficiency at 300 K for one sun under AM1.5 condition is shown in Fig. 44 as a function of the bandgap energy. The slight oscillations are caused by atmospheric absorption. Note that the efficiency has a broad maximum in the E_g range of 0.8 – 1.4 eV. Many factors degrade the ideal efficiency, so that efficiencies actually achieved are lower. Practical solar cells will be considered in subsequent sections. Figure 44 also shows the ideal efficiency at an optical concentration of 1,000 suns (i.e., 844 kW/m^2). Details on optical concentration will be considered in Section 13.9.4. The ideal peak efficiency increases from 31% for 1 sun to 37% for 1,000 suns. This increase is primarily due to the increase of V_{oc}, while the photocurrent is increased linearly with the intensity.

Nonideal Effects. For a practical solar cell, the ideal equivalent circuit, Fig. 41, will be modified to include the series resistance R_s from ohmic loss in the front surface

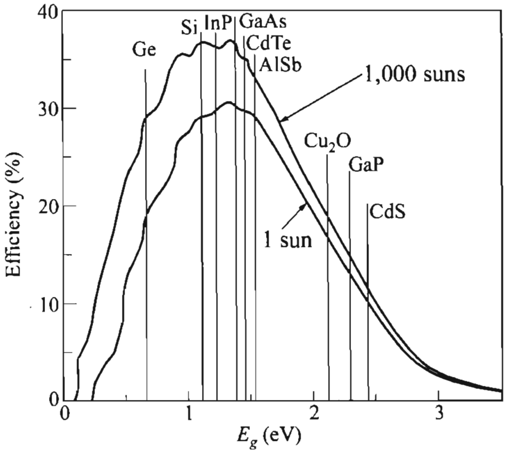


Fig. 44 Ideal solar-cell efficiency at 300 K for 1-sun and 1,000-sun concentration. (After Ref. 82.)

and the shunt resistance R_{sh} from leakage currents. The equivalent circuit should include R_s added in series with the load R_L, and R_{sh} added in parallel with the diode. The diode I-V characteristics are found to be modified from Eq. 106 to[83]

$$\ln\left(\frac{I+I_L}{I_s} - \frac{V-IR_s}{I_s R_{sh}} + 1\right) = \frac{q}{kT}(V-IR_s) . \tag{115}$$

In practice, the shunt resistance has much less effect than the series resistance. The effect of R_s can be simply obtained by replacing V with $(V - IR_s)$, and the main impact is on the fill factor.

For a practical solar cell, the forward current can be dominated by the recombination current in the depletion region. The efficiency is reduced compared with that of an ideal diode. The recombination current is in the form

$$I_{re} = I_s'\left[\exp\left(\frac{qV}{2kT}\right) - 1\right] . \tag{116}$$

The energy conversion equation can again be put into closed form yielding equations similar to Eqs. 107 through 112, with the exception that I_s is replaced by I_s', and the exponential factor is divided by 2. The efficiency for the case of recombination current is found to be much less than the ideal-current case due to degradation of both V_{oc} and the fill factor. For solar cells having mixtures of diffusion current and recombination current, or currents due to other defects, the forward current shows an exponential dependence on the forward voltage as $\exp(qV/nkT)$, where n is called the ideality factor and is generally between 1–2. The efficiency decreases with increasing values of n.

As the device temperature increases, the diffusion lengths will increase because the diffusion constant stays the same or increases with temperature, and the minority-carrier lifetime increases with temperature. The increase in minority-carrier diffusion length will increase the photocurrent I_L. However, V_{oc} will rapidly decrease because of the exponential dependence of the saturation current on temperature. The degradation in the *softness* in the knee of the I-V curve as temperature increases will also decrease the fill factor. Therefore, the overall effect causes a reduction of efficiency as the temperature increases. This presents a challenge for operation under optical concentrators.

For satellite applications, high-energy particle radiation in outer space produces defects in semiconductors that cause a reduction in solar-cell power output. This is due to the decrease of minority-carrier diffusion length caused by bombardment of these high-energy particles. To improve the radiation tolerance, lithium has been incorporated into the solar cells. The Li can diffuse to and combine with radiation-induced point defects.

13.9.3 Photocurrent and Spectral Response

In this section we derive the photocurrent for a silicon p-n junction solar cell which serves as a reference device for all solar cells. A typical schematic representative of a solar cell is shown in Fig. 45. It consists of a shallow p-n junction formed on the sur-

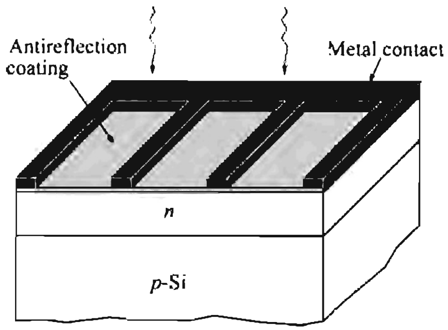


Fig. 45 Schematic representation of a silicon *p-n* junction solar cell.

face, front ohmic contact stripes and fingers, an antireflection coating, and a back ohmic contact. The finger grid reduces series resistance, but at the expense of blocking some light, so there is a trade-off for the design. Some use transparent conductors such as ITO (indium-tin oxide).

When a monochromatic light of wavelength λ is incident on the front surface, the photocurrent and spectral response, that is, the number of carriers collected per incident photon at each wavelength, can be derived as follows. The generation rate of electron-hole pairs at a distance x from the semiconductor surface is shown in Fig. 46 and is given by

$$G(\lambda, x) = \alpha(\lambda)\phi(\lambda)[1-R(\lambda)]\exp[-\alpha(\lambda)x] \tag{117}$$

where $\alpha(\lambda)$ is the absorption coefficient, $\phi(\lambda)$ the number of incident photons per area per time per unit bandwidth, and $R(\lambda)$ the fraction of these photons reflected from the surface.

For an abrupt *p-n* junction solar cell with constant doping on each side, Fig. 46, there are no electric fields outside the depletion region. Photogenerated carriers in these regions are collected by a diffusion process while that in the depletion region by drift process. We divide the collection of photogenerated carriers in three regions: the top neutral region, the depletion region of the junction, and the substrate neutral region. We also assume an abrupt one-sided junction with $N_D \gg N_A$, so the depletion region at the *n*-side can be neglected.

Under low-injection condition, the one-dimensional, steady-state continuity equations are

$$G_n - \left(\frac{n_p - n_{po}}{\tau_n}\right) + \frac{1}{q}\frac{dJ_n}{dx} = 0 \tag{118a}$$

for electrons in the *p*-type substrate and

$$G_p - \left(\frac{p_n - p_{no}}{\tau_p}\right) - \frac{1}{q}\frac{dJ_p}{dx} = 0 \tag{118b}$$

for holes in the *n*-type layer. The current-density equations are

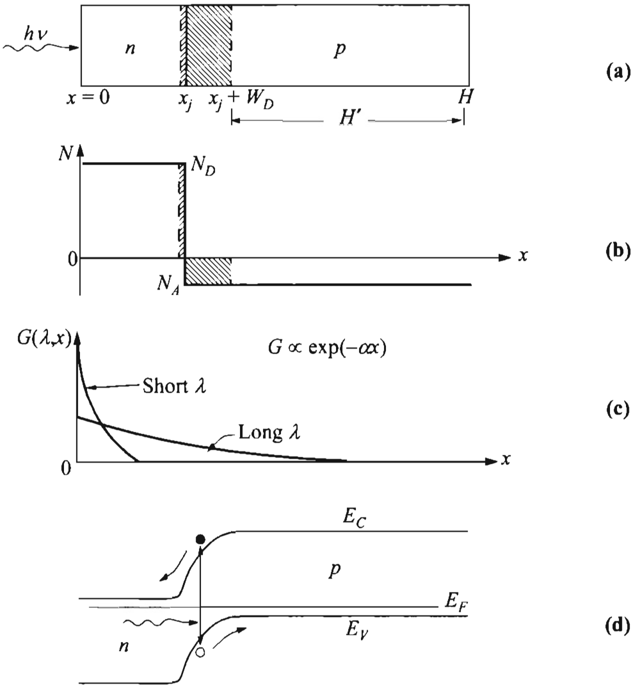


Fig. 46 (a) Solar-cell dimensions under consideration. (b) Assumed abrupt doping profiles $N_D \gg N_A$. (c) Generation rate as a function of distance for long and short wavelengths. (d) Energy-band diagram showing generated electron-hole pair.

$$J_n = q\mu_n n_p \mathscr{E} + qD_n\left(\frac{dn_p}{dx}\right), \tag{119a}$$

$$J_p = q\mu_p p_n \mathscr{E} - qD_p\left(\frac{dp_n}{dx}\right). \tag{119b}$$

For the top n-side of the junction, Eqs. 117, 118b, and 119b can be combined to yield an expression:

$$D_p\frac{d^2p_n}{dx^2} + \alpha\phi(1-R)\exp(-\alpha x) - \frac{p_n - p_{no}}{\tau_p} = 0. \tag{120}$$

The general solution to this equation is

$$p_n - p_{no} = C_2\cosh\left(\frac{x}{L_p}\right) + C_3\sinh\left(\frac{x}{L_p}\right) - \frac{\alpha\phi(1-R)\tau_p}{\alpha^2L_p^2 - 1}\exp(-\alpha x) \tag{121}$$

where $L_p = \sqrt{D_p\tau_p}$ is the diffusion length, and C_2 and C_3 are constants. There are two boundary conditions. At the surface, we have surface recombination with a recombination velocity S_p:

$$D_p \frac{d(p_n - p_{no})}{dx} = S_p(p_n - p_{no}) \qquad \text{at } x = 0 . \tag{122}$$

At the depletion edge, the excess carrier density is small due to the electric field in the depletion region:

$$p_n - p_{no} \approx 0 \qquad \text{at } x = x_j . \tag{123}$$

Using these boundary conditions in Eq. 121, the hole density is

$$p_n - p_{no} = [\alpha\phi(1-R)\tau_p/(\alpha^2L_p^2 - 1)] \times \left[\frac{\left(\frac{S_pL_p}{D_p} + \alpha L_p\right)\sinh\frac{x_j - x}{L_p} + \exp(-\alpha x_j)\left(\frac{S_pL_p}{D_p}\sinh\frac{x}{L_p} + \cosh\frac{x}{L_p}\right)}{(S_pL_p/D_p)\sinh(x_j/L_p) + \cosh(x_j/L_p)} - \exp(-\alpha x) \right] \tag{124}$$

and the resulting hole photocurrent density at the depletion edge is

$$J_p = -qD_p\left(\frac{dp_n}{dx}\right)_{x_j} = [q\phi(1-R)\alpha L_p/(\alpha^2L_p^2 - 1)] \times \left[\frac{\left(\frac{S_pL_p}{D_p} + \alpha L_p\right) - \exp(-\alpha x_j)\left(\frac{S_pL_p}{D_p}\cosh\frac{x_j}{L_p} + \sinh\frac{x_j}{L_p}\right)}{(S_pL_p/D_p)\sin(x_j/L_p) + \cosh(x_j/L_p)} - \alpha L_p \exp(-\alpha x_j) \right] . \tag{125}$$

This photocurrent would be generated and collected in the front side of an *n*-on-*p* junction solar cell at a given wavelength, assuming this region to be uniform in lifetime, mobility, and doping level.

To find the electron photocurrent generated in the substrate of the cell, Eqs. 117, 118a, and 119a are used with the boundary conditions:

$$n_p - n_{po} \approx 0 \qquad \text{at } x = x_j + W_D \tag{126}$$

$$S_n(n_p - n_{po}) = \frac{-D_n dn_p}{dx} \qquad \text{at } x = H \tag{127}$$

where W_D is the depletion width and H is the width of the entire cell. Equation 126 states that the excess minority carrier density is near zero at the edge of the depletion region, while Eq. 127 states that the back surface recombination takes place at the ohmic contact.

Using these boundary conditions, the electron distribution in a uniformly doped *p*-type substrate is

$$n_p - n_{po} = \frac{\alpha\phi(1-R)\tau_n}{\alpha^2L_n^2 - 1}\exp[-\alpha(x_j + W_D)]\left\{ \cosh\left(\frac{x'}{L_n}\right) - \exp(-\alpha x') - \frac{(S_nL_n/D_n)[\cosh(H'/L_n) - \exp(-\alpha H')] + \sinh(H'/L_n) + \alpha L_n \exp(-\alpha H')}{(S_nL_n/D_n)\sinh(H'/L_n) + \cosh(H'/L_n)} \times \sinh(x'/L_n) \right\} \tag{128}$$

($x' \equiv x - x_j - W_D$) and the photocurrent due to electrons collected at the depletion edge, $x = x_j + W_D$, is

$$J_n = qD_n\left(\frac{dn_p}{dx}\right)_{x_j+W_D} = \frac{q\phi(1-R)\alpha L_n}{\alpha^2L_n^2-1}\exp[-\alpha(x_j+W_D)] \times \left\{\alpha L_n - \frac{(S_nL_n/D_n)[\cosh(H'/L_n)-\exp(-\alpha H')]+\sinh(H'/L_n)+\alpha L_n\exp(-\alpha H')}{(S_nL_n/D_n)\sinh(H'/L_n)+\cosh(H'/L_n)}\right\} \tag{129}$$

where H' as shown in Fig. 46a is the p-substrate neutral region.

Some photocurrent generation takes place within the depletion region as well. The electric field in this region is generally high, and the photogenerated carriers are accelerated out of the depletion region before they can recombine. The quantum efficiency in this region is near 100% and the photocurrent per unit bandwidth is equal to the number of photons absorbed:

$$J_{dr} = q\phi(1-R)\exp(-\alpha x_j)[1-\exp(-\alpha W_D)]. \tag{130}$$

The total photocurrent at a given wavelength is then the sum of Eqs. 125, 129, and 130:

$$J_L(\lambda) = J_p(\lambda)+J_n(\lambda)+J_{dr}(\lambda). \tag{131}$$

The spectral response (SR) is defined as this sum divided by $q\phi$ for externally observed response or by $q\phi(1-R)$ for internal SR:

$$\mathrm{SR}(\lambda) = \frac{J_L(\lambda)}{q\phi(\lambda)[1-R(\lambda)]} = \frac{J_p(\lambda)+J_n(\lambda)+J_{dr}(\lambda)}{q\phi(\lambda)[1-R(\lambda)]}. \tag{132}$$

The ideal internal SR for a semiconductor with energy gap E_g is a step function that equals zero for $h\nu < E_g$ and unity for $h\nu \geq E_g$ (dashed line in Fig. 47a). A realistic internal SR calculated for a Si n-p solar cell is shown in Fig. 47a, which departs substantially from the idealized step function at high photon energies. The figure also

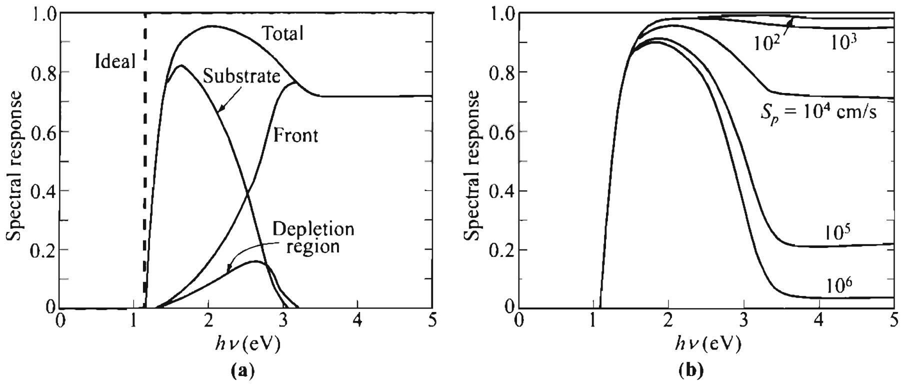


Fig. 47 (a) Computed internal spectral response of Si n-on-p cell, showing the individual contributions from each of the three regions. Dashed curve is for ideal response. Parameters used: $N_D = 5\times10^{19}$ cm^{-3}, $N_A = 1.5\times10^{16}$ cm^{-3}, $\tau_p = 0.4$ μs, $\tau_n = 10$ μs, $x_j = 0.5$ μm, $H = 450$ μm, S_p(front) $= 10^4$ cm/s, and S_n(back) $= \infty$. (b) Computed internal spectral response with different surface recombination velocities. (After Ref. 84.)

shows the individual contributions from each of the three regions. At low photon energies, most carriers are generated in the substrate region because of the low absorption coefficient in Si. As the photon energy increases above 2.5 eV, the front region takes over. Above 3.5 eV, α becomes larger than 10^6 cm^{-1}, and the SR derives entirely from the front region. Since S_p is assumed to be quite high, the surface recombination at the front region causes large departure from the ideal response. The SR approaches an asymptotic value when $\alpha L_p \gg 1$ and $\alpha x_j \gg 1$ (i.e., from the front-side photocurrent, Eq. 125):

$$\mathrm{SR} = \frac{1 + (S_p/\alpha D_p)}{(S_p L_p/D_p)\sinh(x_j/L_p) + \cosh(x_j/L_p)} . \tag{133}$$

The surface recombination velocity S_p has a profound effect on the SR especially at high photon energies. This effect is illustrated in Fig. 47b for devices with the same parameters as in Fig. 47a, except that S_p varies from 10^2 to 10^6 cm/s. Note the drastic reduction in SR as S_p increases. Equation 133 also shows that for a given S_p, the SR can be improved by increasing the diffusion length L_p. Generally, to increase SR over the useful wavelength range, one should increase both L_n and L_p, and decrease both S_n and S_p.

Once the SR is known, the total photocurrent density obtained from the solar spectral distribution $\phi(\lambda)$, shown in Fig. 39, is given by

$$J_L = q \int_0^{\lambda_m} \phi(\lambda)[1 - R(\lambda)]\mathrm{SR}(\lambda) d\lambda \tag{134}$$

where λ_m is the longest wavelength corresponding to the semiconductor bandgap. To obtain a large J_L, one should minimize $R(\lambda)$ and maximize $\mathrm{SR}(\lambda)$ over the wavelength range $0 < \lambda < \lambda_m$.

13.9.4 Device Configuration

The main requirements for the solar cells are high efficiency, low cost, and good reliability. Many solar-cell configurations have been proposed and demonstrated with impressive success. However, in order for solar cells to have an impact on production of the total energy consumption, more challenge is still ahead but the goal is achievable to many believers. We shall consider a few main solar-cell designs and their device performances.

Crystal-Si Solar Cell. The crystal-Si solar cell is enjoying the most success in the market at the present time. It has a reasonable balance between performance and cost. Its best reported efficiency has reached higher than 22%. The main cost is the crystal substrate, and a great deal of research is focused on reducing the cost of the crystal growth. The requirement of the crystal quality is less stringent than that for high-density integrated circuits. One approach is the ribbon growth technique from the Si melt. Instead of the regular ingot shape, the crystal is pulled in a thin sheet with a thickness smaller than a typical Si wafer. This technique reduces cost in both the process of cutting the ingot into wafers and material waste during the cutting. We will

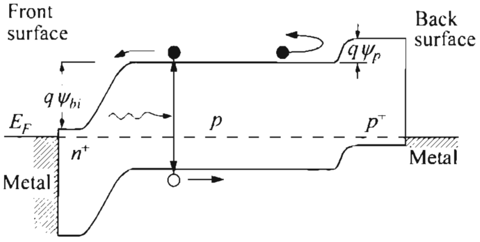


Fig. 48 Energy-band diagram for an n^+-p-p^+ back-surface field junction solar cell. (After Ref. 85.)

consider in the following other features that have contributed to the high performance of the solar cell.

The idea of a *back surface field* (BSF) has improved the output voltage over that in conventional cells. A schematic band diagram is shown in Fig. 48. The front surface is made in the normal way, but the back of the cell has a very heavily doped region adjacent to the contact. The potential energy barrier $q\psi_p$ tends to confine minority carriers (electrons) in the more lightly doped region and helps to drive them toward the front. The BSF cell is equivalent to a normal cell having a very small recombination velocity at the back ($S_n < 100$ cm/s). The low S_n will enhance the spectral response at low photon energies. Therefore, the short-circuit current density will increase. The open-circuit voltage is also increased due to increased short-circuit current, decreased diode recombination current at the back contact, and the added potential energy $q\psi_p$.

To reduce light reflection, a textured surface, either on the front or on the back, has been used to trap the light. The *textured* cell, as an example, has pyramidal surfaces produced by anisotropic etching of ⟨100⟩-oriented Si surface as shown in Fig. 49. Light incident on the side of a pyramid will be reflected onto another pyramid instead of being reflected backward. The reflectivity of bare Si is reduced from about 35% for flat surfaces to around 20% for the textured surface. The addition of an antireflection coating reduces the overall reflection to a few percent.

Another area of cost saving is the thick metallization process since solar cells are power devices and they conduct much higher current than regular integrated circuits. A process known as screen printing is commonly used to deposit thick metal layers in

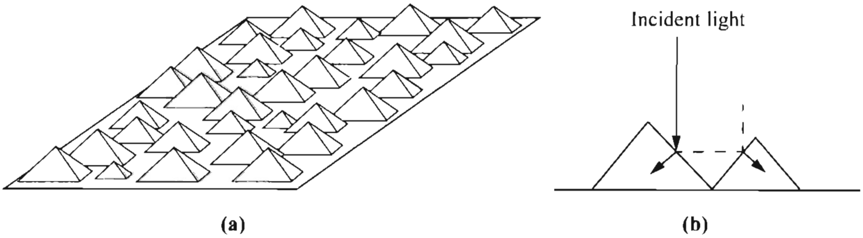


Fig. 49 (a) Textured cell with pyramidal surfaces. (b) Optical path showing the trapping of light to reduce reflection. (After Ref. 86.)

a production environment. This process is much faster than metal deposition in vacuum systems.

Thin-Film Solar Cells. In thin-film solar cells, the active semiconductor layers are polycrystalline or disordered films that have been deposited or formed on electrically active or passive substrates, such as glass, plastic, ceramic, metal, graphite, or metallurgical silicon. A thin film of semiconductor can be deposited onto a foreign substrate by various methods, such as vapor growth, plasma evaporation, and plating. If the semiconductor thickness is larger than the absorption length, most light will be absorbed; and if the diffusion length is larger than the film thickness, most photogenerated carriers can be collected. The most-common and successful films used are Si, CdTe, CdS, CIS ($CuInSe_2$), and CIGS ($CuInGaSe_2$). Their efficiencies have reached higher than 15%.

The main advantage of thin-film solar cells is their promise of low cost, due to low-cost processing and the use of relatively low-cost materials. The main disadvantages are low efficiency and long-term instability. The low efficiency is partly caused by the grain-boundary effects and the poor quality of the semiconductor material grown on foreign substrates. Another problem is poor stability which is caused by the chemical reaction of the semiconductor with the ambient such as O_2 and water vapor. Steps must be taken to ensure device reliability in thin-film solar cells.

Amorphous silicon (a-Si) is a well-studied material for thin-film solar cells. Layers 1- to 3-μm thick are grown by RF glow-discharge decomposition of silane onto metal or glass substrates. The difference between crystalline and amorphous Si is dramatic; the former has an indirect bandgap of 1.1 eV, whereas hydrogenated a-Si has an optical absorption characteristic resembles that expected for a crystal with a direct bandgap of 1.6 eV. Solar cells have been prepared in *p-n* junction as well as Schottky-barrier. Since the absorption coefficient is in the range 10^4 to 10^5 cm^{-1} across the visible portion of the solar spectrum, many carriers are photogenerated within a fraction of a micron of the illuminated surface.

Since deposited thin films invariably contain traps, we now estimate at what concentration level the traps will start to severely degrade the device performance. In the absence of charged traps, the electric field will be uniform and is given by $\mathscr{E} = E_g/qH$ where H is the total film thickness. For a thickness of $1/\alpha$ ($\approx$ 0.1 μm) and $E_g = 1.5$ eV, the field is 1.5×10^5 V/cm. When traps are present with a concentration n_t the net space charge will be n_c, where $n_c < n_t$. These charged defects will affect the field strength by $\Delta\mathscr{E} = qn_cH/\varepsilon_s$. Assuming a dielectric constant of 4, we find that $\Delta\mathscr{E} \ll \mathscr{E}$ if $n_c < 10^{16}$ cm^{-3}, implying that a total trap concentration as high as 10^{17} cm^{-3} can be tolerated without serious disturbance to the electric field. A further requirement is that the space-charge-limited current must be higher than about 100 mA/cm^2, substantially larger than the short-circuit current density produced in one-sun illumination. For 0.1-μm thickness, this condition also leads to a permissible trap density as high as 10^{17} cm^{-2}/eV.

The electric field must also be able to extract the electrons and holes in a transit time $H/\mathscr{E}\mu$, which is short compared with the recombination lifetime $(n_t v\sigma)^{-1}$, where

σ is the capture cross section ($\approx 10^{-14}$ cm^2), and υ is the thermal velocity ($\approx 10^7$ cm/s). This condition will be satisfied if

$$\mu > \frac{n_t \upsilon \sigma H}{\mathscr{E}} = \frac{n_t \upsilon \sigma q H^2}{E_g} \approx 1 \text{ cm}^2/\text{V-s} \tag{135}$$

which is not difficult to achieve.

The considerations discussed above imply that useful solar cells can be made in semiconductors containing very high defect density if the semiconductor films are sufficiently thin and have high absorption coefficient near the band edge, coupled with requisite mobilities.

Schottky-Barrier and MIS Solar Cells. The basic characteristics of Schottky-barrier diodes have been described in Chapter 3. The metal must be thin enough to allow a substantial amount of light to reach the semiconductor. Short-wavelength light entering the semiconductor is mainly absorbed in the depletion region. Long-wavelength light is absorbed in the neutral region, creating electron-hole pairs just as in a *p-n* junction. For solar-cell applications, the excitation of carriers from the metal into the semiconductor contributes less than 1% to the total photocurrent and, therefore, can be neglected.

The advantages of Schottky barriers include (1) low-temperature processing because no high-temperature diffusion or annealing is required; (2) adaptability to polycrystalline and thin-film solar cells; (3) high radiation resistance due to high electrical field near the surface; and (4) high-current output and good spectral response, because the presence of a depletion region right at the semiconductor surface can substantially reduce the effects of low lifetime and high recombination velocity near the surface.

The two major contributions to the photocurrent come from the depletion region and the substrate neutral region. The collection from the depletion region is similar to that of a *p-n* junction, leading to a photocurrent:

$$J_{dr} = qT(\lambda)\phi(\lambda)[1 - \exp(-\alpha W_D)] \tag{136}$$

where $T(\lambda)$ is the transmission coefficient of the metal. The photocurrent from the substrate region is given by an expression identical to Eq. 129, except that $(1 - R)$ is replaced by $T(\lambda)$, and $\alpha(x_j + W_D)$ by αW_D. If the back contact is ohmic and the device thickness is much greater than the diffusion length $H' \gg L_p$, the photocurrent from the substrate region is simplified to

$$J_n = qT(\lambda)\phi(\lambda)\frac{\alpha L_n}{\alpha L_n + 1}\exp(-\alpha W_D). \tag{137}$$

The total photocurrent is given by the sum of Eqs. 136 and 137.

The *I-V* characteristics of a Schottky barrier under illumination is given by

$$I = I_s\left[\exp\left(\frac{qV}{nkT}\right) - 1\right] - I_L \tag{138}$$

and

$$I_s = AA^{**}T^2\exp\left(\frac{-q\phi_B}{kT}\right) \tag{139}$$

where n is the ideality factor, A^{**} the effective Richardson constant (refer to Chapter 3), and $q\phi_B$ the barrier height. The conversion efficiency is given by Eq. 114. For a given semiconductor, the efficiency can be calculated from Eqs. 114, 137, and 138 as a function of the barrier height.

For most metal-semiconductor systems made on uniformly doped substrates, the maximum barrier height is about $(2/3)E_g$. Consequently the built-in potential is lower than that of a *p-n* junction, so the V_{oc} is also lower. However, the barrier height can be increased to near the bandgap energy by inserting a thin, heavily doped layer (10 nm) of opposite type near the semiconductor surface.

In an MIS (metal-insulator-semiconductor) solar cell, a thin insulating layer is inserted between the metal and semiconductor surface. The advantages of MIS solar cells include an electric field extending to the semiconductor surface in a direction that aids in collecting minority carriers generated by short-wavelength light, and the fact that the active region of the cells is free of the diffusion-induced crystal damage inherent in diffused *p-n* junction cells. The saturation current density is similar to that for Schottky barrier with an additional tunneling term (refer to Chapter 8):

$$J_s = A^{**}T^2\exp\left(\frac{-q\phi_B}{kT}\right)\exp(-\delta\sqrt{q\phi_T}) \tag{140}$$

where $q\phi_T$ in eV is the average barrier height presented by the insulating layer and δ in Å is the insulator thickness. Substitution of $V = V_{oc}$ and $J = 0$ in Eq. 138 yields,

$$V_{oc} = \frac{nkT}{q}\left[\ln\left(\frac{J_L}{A^{**}T^2}\right) + \frac{q\phi_B}{kT} + \delta\sqrt{q\phi_T}\right]. \tag{141}$$

Equation 141 shows that V_{oc} of an MIS solar cell will increase with increasing δ. However, as the insulator thickness δ increases further, the short-circuit current will decrease, causing a degradation of the conversion efficiency. An optimum oxide thickness for an MIS cell is found to be about 2 nm.[87]

Multiple-Junction Solar Cell. The theoretical maximum efficiency discussed is based on the balance of E_g for both photocurrent and open-circuit voltage. It has been shown that when using multiple junctions of different bandgaps and stacked on top of each other, the efficiency can be improved because there is less waste of photons below a single E_g. A theoretical calculation on two-junction solar cell is shown in Fig. 50. For this tandem cell, the maximum efficiency is ≈ 40% with E_{g1} = 1.7 eV and E_{g2} = 1 eV. For three-junction cells, the ideal combination is E_{g1} = 1.75 eV, E_{g2} = 1.18 eV, and E_{g3} = 0.75 eV. Beyond the three bandgaps, the efficiency increases very slightly. Experimentally, on crystal solar cells, the three-junction cells based on compound semiconductors of GaAs/InGaAs and InGaP/InGaAs/Ge have shown efficiencies of higher than 30%, the highest ever report of any structure. On thin-film cells, the multiple-junction cells on SiGeC/Si/SiGe and SiGeC/Si/GeC have been demonstrated to have higher efficiencies than their single-junction counterparts.

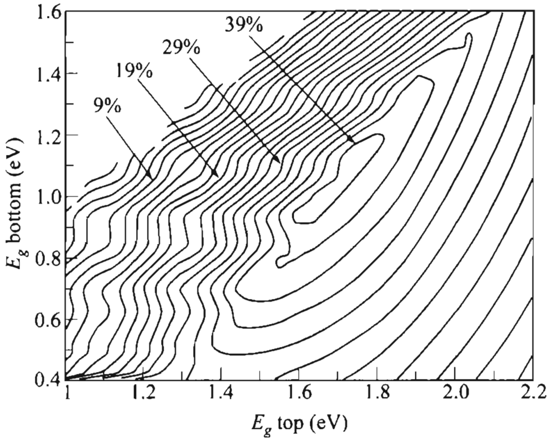


Fig. 50 Maximum efficiency of tandem solar cell as a function of the top E_g and bottom E_g. (After Ref. 88.)

Optical Concentration. Sunlight can be focused by using mirrors and lenses. Optical concentration offers an attractive and flexible approach to reducing high cell costs by substituting a concentrator area for much of the cell area. It also offers other advantages, including (1) increased cell efficiency (Fig. 44), (2) hybrid systems yielding both electrical and thermal outputs, and (3) reduced cell-temperature coefficient.

In a standard concentrator module, mirrors and lens are used to direct and focus the sunlight onto the solar cells mounted on a water-cooled block. The experimental results obtained from a silicon vertical-junction solar cell are shown in Fig. 51. Note that device performances improve as the concentration increases from 1 sun toward

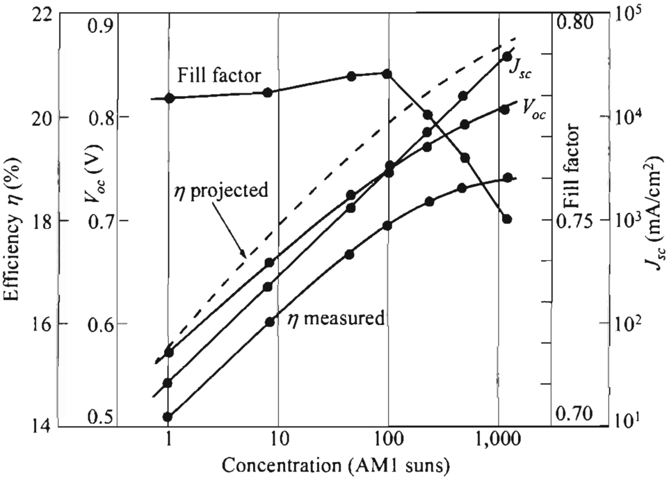


Fig. 51 Efficiency η, V_{oc}, J_{sc}, and fill factor vs. AM1 solar concentration for a multiple vertical junction solar cell. (After Ref. 89.)

1,000 suns. The short-circuit current increases linearly with concentration. The open-circuit voltage increases at a rate of 0.1 V per decade of intensity, while the fill factor degrades slightly. The efficiency, which is the product of the foregoing three factors divided by the input concentrated power, increases at a rate of about 2% per decade. Therefore, one cell operated under 1,000-sun concentration can produce the same power output as 1,300 cells under 1 sun. Hence the optical concentration approach can potentially replace expensive solar cells with less expensive concentrator materials and a related tracking setup to minimize the overall system cost.

Under high concentrations, the carrier density approaches that of the substrate doping, and a high injection condition prevails. The current density is proportional to $\exp(qV/nkT)$, where $n = 2$. The open-circuit voltage becomes

$$V_{oc} = \frac{2kT}{q}\ln\left(\frac{J_L}{J_s} + 1\right) \tag{142}$$

and I_s can be expressed as

$$J_s = C_4\left(\frac{T}{T_0}\right)^{3/2}\exp\left[-\frac{E_g(T)}{2kT}\right] \tag{143}$$

where C_4 is a constant, T is the operating temperature, and T_0 is 300 K. The temperature coefficient of V_{oc} is found to change from –2.07 mV/°C at 1 sun to –1.45 mV/°C at 500 suns. Thus for silicon solar cells, high solar concentration levels can decrease the efficiency from losses associated with operation at elevated temperatures.

The efficiency and power output can be increased further by using individual solar cells of different bandgaps. A spectral splitting arrangement diverges the solar light flux into many narrow spectral bands and delivers the flux in each band to a cell with the bandgap value optimized for that band.

REFERENCES

1. H. Melchior, "Demodulation and Photodetection Techniques," in F. T. Arecchi and E. O. Schulz-Dubois, Eds., *Laser Handbook*, Vol. **1**, North-Holland, Amsterdam, 1972, pp. 725–835.
2. M. Ross, *Laser Receivers-Devices, Techniques, Systems*, Wiley, New York, 1966.
3. C. A. Musca, J. F. Siliquini, B. D. Nener, and L. Faraone, "Heterojunction Blocking Contacts in MOCVD Grown $Hg_{1-x}Cd_xTe$ Long Wavelength Infrared Photoconductors," *IEEE Trans. Electron Dev.*, **ED-44**, 239 (1997).
4. M. DiDomenico, Jr. and O. Svelto, "Solid State Photodetection Comparison between Photodiodes and Photoconductors," *Proc. IEEE*, **52**, 136 (1964).
5. A. Van der Ziel, *Fluctuation Phenomena in Semiconductors*, Academic, New York, 1959, Chap. 6.
6. W. L. Eisenman, J. D. Merriam, and R. F. Potter, "Operational Characteristics of Infrared Photodiode," in R. K. Willardson and A. C. Bear, Eds., *Semiconductors and Semimetals*, Vol. **12**, *Infrared Detector II*, Academic, New York, 1977, pp. 1–38.

7. G. E. Stillman, C. M. Wolfe, and J. O. Dimmock, "Far-Infrared Photoconductivity in High Purity GaAs," in R. K. Willardson and A. C. Bear, Eds., *Semiconductors and Semimetals*, Vol. **12**, *Infrared Detector II*, Academic, New York, 1977, pp. 169–290.
8. G. E. Stillman and C. M. Wolfe, "Avalanche Photodiode," in R. K. Willardson and A. C. Bear, Eds., *Semiconductors and Semimetals*, Vol. **12**, *Infrared Detector II*, Academic, New York, 1977, pp. 291–394.
9. R. G. Smith and S. D. Personick, "Receiver Design for Optical Communication Systems," in H. Kressel, Ed., *Semiconductor Devices for Optical Communication*, Springer-Verlag, New York, 1979, Chap. 4.
10. W. W. Gartner, "Depletion-Layer Photoeffects in Semiconductors," *Phys. Rev.*, **116**, 84 (1959).
11. H. S. Lee and S. M. Sze, "Silicon *p-i-n* Photodetector Using Internal Reflection Method," *IEEE Trans. Electron Dev.*, **ED-17**, 342 (1970).
12. J. Muller, "Thin Silicon Film *p-i-n* Photodiodes with Internal Reflection," *IEEE Trans. Electron Dev.*, **ED-25**, 247 (1978).
13. K. Ahmad and A. W. Mabbitt, "GaInAs Photodiodes," *Solid-State Electron.*, **22**, 327 (1979).
14. W. F. Kosonocky, "Review of Schottky-Barrier Imager Technology," *SPIE*, **1308**, 2 (1990).
15. C. R. Crowell and S. M. Sze, "Hot Electron Transport and Electron Tunneling in Thin Film Structures," in R. E. Thun, Ed., *Physics of Thin Films*, Vol. **4**, Academic, New York, 1967, pp. 325–371.
16. W. M. Sharpless, "Cartridge-Type Point Contact Photodiode," *Proc. IEEE,*, **52**, 207 (1964).
17. J. C. Campbell, S. Demiguel, F. Ma, A. Beck, X. Guo, S. Wang, X. Zheng, X. Li, J. D. Beck, M. A. Kinch, A. Huntington, L. A. Coldren, J. Decobert, and N. Tscherptner, "Recent Advances in Avalanche Photodiodes," *IEEE J. Selected Topics Quan. Elect.*, **10**, 777 (2004).
18. H. Melchior and W. T. Lynch, "Signal and Noise Response of High Speed Germanium Avalanche Photodiodes," *IEEE Trans. Electron Dev.*, **ED-13**, 829 (1966).
19. R. B. Emmons, "Avalanche Photodiode Frequency Response," *J. Appl. Phys.*, **38**, 3705 (1967).
20. R. J. McIntyre, "Multiplication Noise in Uniform Avalanche Diodes," *IEEE Trans. Electron Dev.*, **ED-13**, 164 (1966).
21. R. D. Baertsch, "Noise and Ionization Rate Measurements in Silicon Photodiodes," *IEEE Trans. Electron Dev.*, **ED-13**, 987 (1966).
22. R. J. McIntyre, "The Distribution of Gains in Uniformly Multiplying Avalanche Photodiodes: Theory," *IEEE Trans. Electron Dev.*, **ED-19**, 703 (1972).
23. R. P. Webb, R. J. McIntyre, and J. Conradi, "Properties of Avalanche Photodiodes," *RCA Rev.*, **35**, 234 (1974).
24. H. Kanbe and T. Kmura, "Figure of Merit for Avalanche Photodiodes," *Electron. Lett.*, **13**, 262 (1977).
25. L. K. Anderson, P. G. McMullin, L. A. D'Asaro, and A. Goetzberger, "Microwave Photodiodes Exhibiting Microplasma-Free Carrier Multiplication," *Appl. Phys. Lett.*, **6**, 62 (1965).

26. S. M. Sze and G. Gibbons, "Effect of Junction Curvature on Breakdown Voltage in Semiconductors," *Solid-State Electron.*, **9**, 831 (1966).
27. H. W. Ruegg, "An Optimized Avalanche Photodiode," *IEEE Trans. Electron Dev.*, **ED-14**, 239 (1967).
28. J. Moll, *Physics of Semiconductors*, McGraw-Hill, New York, 1964.
29. H. Melchior, A. R. Hartman, D. P. Schinke, and T. E. Seidel, "Planar Epitaxial Silicon Avalanche Photodiode," *Bell Syst. Tech. J.*, **57**, 1791 (1978).
30. H. Melchior, M. P. Lepselter, and S. M. Sze, "Metal-Semiconductor Avalanche Photodiode," *IEEE Solid-State Device Res. Conf.*, Boulder, Colo., June 17–19, 1968.
31. N. Susa, H. Nakagome, O. Mikami, H. Ando, and H. Kanbe, "New InGaAs/InP Avalanche Photodiode Structure for the 1–1.6 μm Wavelength Region" *IEEE J. Quantum Electron.*, **QE-16**, 864 (1980).
32. O. Hildebrand, W. Kuebart, and M. H. Pilkuhn, "Resonant Enhancement of Impact Ionization in $Al_xGa_{1-x}Sb$ *p-i-n* Avalanche Photodiodes," *Appl. Phys. Lett.*, **37**, 801 (1980).
33. F. H. DeLaMoneda, E. R. Chenette, and A. Van der Ziel, "Noise in Phototransistors," *IEEE Trans. Electron Dev.*, **ED-18**, 340 (1971).
34. S. Knight, L. R. Dawson, U. G. Keramidas, and M. G. Spencer, "An Optically Triggered Double Heterostructure Linear Bilateral Phototransistor," *Tech. Dig. IEEE IEDM*, 1977, p. 472.
35. W. S. Boyle and G. E. Smith, "Charge Coupled Semiconductor Devices," *Bell Syst. Tech. J.*, **49**, 587 (1970).
36. M. F. Tompsett, G. F. Amelio, and G. E. Smith, "Charge Coupled 8-bit Shift Register," *Appl. Phys. Lett.*, **17**, 111 (1970).
37. M. F. Tompsett, G. F. Amelio, W. J. Bertram, Jr., R. R. Buckley, W. J. McNamara, J. C. Mikkelsen, Jr., and D. A. Sealer, "Charge-Coupled Imaging Devices: Experimental Results," *IEEE Trans. Electron Dev.*, **ED-18**, 992 (1971).
38. W. J. Bertram, D. A. Sealer, C. H. Sequin, M. F. Tompsett, and R. R. Buckley, "Recent Advances in Charge Coupled Imaging Devices," *INTERCON Dig.*, 292 (1972).
39. T. F. Tao, J. R. Ellis, L. Kost, and A. Doshier, "Feasibility Study of PbTe and $Pb_{0.76}Sn_{0.24}Te$ Infrared Charge Coupled Imager," *Proc. Int. Conf. Tech. Appl. Charge Coupled Devices*, 259 (1973).
40. J. C. Kim, "InSb MIS Structures for Infrared Imaging Devices," *Tech. Dig. IEEE IEDM*, 419 (1973).
41. H. K. Burke and G. J. Michon, "Charge-Injection Imaging: Operating Techniques and Performances Characteristics," *IEEE Trans. Electron Dev.*, **ED-23**, 189 (1976).
42. W. S. Boyle and G. E. Smith, "Charge-Coupled Devices—A New Approach to MIS Device Structures," *IEEE Spectrum*, **8**, 18 (1971).
43. F. L. J. Sangster, "Integrated MOS and Bipolar Analog Delay Lines Using Bucket-Brigade Capacitor Storage," *Proc. IEEE Int. Solid-State Circuits Conf.*, 74 (1970).
44. M. F. Tompsett, "Video-Signal Generation," in T. P. McLean and P. Schagen, Eds., *Electronic Imaging*, Academic, New York, 1979, p. 55.
45. C. K. Kim, "The Physics of Charge-Coupled Devices," in M. J. Howes and D. V. Morgan, Eds., *Charge-Coupled Devices and Systems*, Wiley, New York, 1979, p. 1.
46. C. H. Sequin and M. F. Tompsett, *Charge Transfer Devices*, Academic, New York, 1975.

47. J. E. Carnes, W. F. Kosonocky, and E. G. Ramberg, "Free Charge Transfer in Charge-Coupled Devices," *IEEE Trans. Electron Dev.*, **ED-19**, 798 (1972).
48. M. H. Elsaid, S. G. Chamberlain, and L. A. K. Watt, "Computer Model and Charge Transport Studies in Short Gate Charge-Coupled Devices," *Solid-State Electron.*, **20**, 61 (1977).
49. M. F. Tompsett, "The Quantitative Effect of Interface States on the Performance of Charge-Coupled Devices," *IEEE Trans. Electron Dev.*, **ED-20**, 45 (1973).
50. R. E. Colbeth and R. A. LaRue, "A CCD Frequency Prescaler for Broadband Applications," *IEEE J. Solid-St. Circuits*, **28**, 922 (1993).
51. M. F. Tompsett, "Charge Transfer Devices," *J. Vac. Sci. Technol.*, **9**, 1166 (1972).
52. W. S. Boyle and G. E. Smith, U.S. Patent 3,792,322 (1974).
53. R. H. Walden, R. H. Krambeck, R. J. Strain, J. McKenna, N. L. Schryer, and G. E. Smith, "The Buried Channel Charge Coupled Device," *Bell Syst. Tech. J.*, **51**, 1635 (1972).
54. A. El Gamal and H. Eltoukhy, "CMOS Image Sensors," *IEEE Circuits Dev. Mag.*, 6, (May/June 2005).
55. K. K. Ng, *Complete Guide to Semiconductor Devices*, 2nd Ed., Wiley/IEEE Press, Hoboken, New Jersey, 2002.
56. T. Sugeta, T. Urisu, S. Sakata, and Y. Mizushima, "Metal-Semiconductor-Metal Photodetector for High-Speed Optoelectronic Circuits," *Proc. 11th Conf. (1979 Int.) Solid State Devices*, Tokyo, 1979. *Jpn. J. Appl. Phys.*, Suppl. **19-1**, 459 (1980).
57. T. Sugeta and T. Urisu, "High-Gain Metal-Semiconductor-Metal Photodetectors for High-Speed Optoelectronics Circuits," *Proc. IEEE Dev. Research Conf.*, 1979. Also in *IEEE Trans. Electron Dev.*, **ED-26**, 1855 (1979).
58. H. Schumacher, H. P. Leblanc, J. Soole, and R. Bhat, "An Investigation of the Optoelectronic Response of GaAs/InGaAs MSM Photodetectors," *IEEE Electron Dev. Lett.*, **EDL-9**, 607 (1988).
59. J. B. D. Soole, H. Schumacher, R. Esagui, and R. Bhat, "Waveguide Integrated MSM Photodetector for the 1.3μm–1.6μm Wavelength Range," *Tech. Dig. IEEE IEDM*, 483 (1988).
60. S. M. Sze, D. J. Coleman, Jr., and A. Loya, "Current Transport in Metal-Semiconductor-Metal (MSM) Structures," *Solid-State Electron.*, **14**, 1209 (1971).
61. B. J. van Zeghbroeck, W. Patrick, J. Halbout, and P. Vettiger, "105-GHz Bandwidth Metal-Semiconductor-Metal Photodiode," *IEEE Electron Dev. Lett.*, **EDL-9**, 527 (1988).
62. J. Kim, W. B. Johnson, S. Kanakaraju, L. C. Calhoun, and C. H. Lee, "Improvement of Dark Current Using InP/InGaAsP Transition Layer in Large-Area InGaAs MSM Photodetectors," *IEEE Trans. Electron Dev.*, **ED-51**, 351 (2004).
63. L. C. Chiu, J. S. Smith, S. Margalit, A. Yariv, and A. Y. Cho, "Application of Internal Photoemission from Quantum-Well and Heterojunction Superlattices to Infrared Photodetectors," *Infrared Phys.*, **23**, 93 (1983).
64. J. S. Smith, L. C. Chiu, S. Margalit, A. Yariv, and A. Y. Cho, "A New Infrared Detector Using Electron Emission from Multiple Quantum Wells," *J. Vac. Sci. Technol.*, **B1**, 376 (1983).
65. L. C. West and S. J. Eglash, "First Observation of an Extremely Large-Dipole Infrared Transition Within the Conduction Band of a GaAs Quantum Well," *Appl. Phys. Lett.*, **46**, 1156 (1985).

66. B. F. Levine, K. K. Choi, C. G. Bethea, J. Walker, and R. J. Malik, "New 10 μm Infrared Detector Using Intersubband Absorption in Resonant Tunneling GaAlAs Superlattices," *Appl. Phys. Lett.*, **50**, 1092 (1987).
67. K. K. Choi, B. F. Levine, C. G. Bethea, J. Walker, and R. J. Malik, "Multiple Quantum Well 10 μm GaAs/$\text{Al}_x\text{Ga}_{1-x}\text{As}$ Infrared Detector with Improved Responsivity," *Appl. Phys. Lett.*, **50**, 1814 (1987).
68. B. F. Levine, C. G. Bethea, G. Hasnain, J. Walker, and R. J. Malik, "High-detectivity $D^* = 1.0\times10^{10}$ cm-$\text{Hz}^{0.5}$/W GaAs/AlGaAs Multiquantum Well $\lambda = 8.3$ μm Infrared Detector," *Appl. Phys. Lett.*, **53**, 296 (1988).
69. L. S. Yu and S. S. Li, "A Metal Grating Coupled Bound-to-Miniband Transition GaAs Multiquantum Well/Superlattice Infrared Detector," *Appl. Phys. Lett.*, **59**, 1332 (1991).
70. B. F. Levine, A. Zussman, J. M. Kuo, and J. de Jong, "19 μm Cutoff Long-Wavelength GaAs/$\text{Al}_x\text{Ga}_{1-x}\text{As}$ Quantum-Well Infrared Photodetectors," *J. Appl. Phys.*, **71**, 5130 (1992).
71. H. C. Liu, "Photoconductive Gain Mechanism of Quantum-Well Intersubband Infrared Detectors," *Appl. Phys. Lett.*, **60**, 1507 (1992).
72. B. F. Levine, "Quantum-Well Infrared Photodetectors," *J. Appl. Phys.*, **74**, R1 (1993).
73. E. Becquerel, "On Electric Effects under the Influence of Solar Radiation," *Compt. Rend.*, **9**, 561 (1839).
74. R. S. Ohl, "Light-Sensitive Electric Device," U.S. Patent 2,402,662. Filed May 27, 1941. Granted June 25, 1946.
75. M. Riordan and L. Hoddeson, "The Origins of the *pn* Junction," *IEEE Spectrum*, **34**, 46 (1997).
76. S. Benzer, "Excess-Defect Germanium Contacts," *Phys. Rev.*, **72**, 1267 (1947).
77. J. I. Pantchechnikoff, "A Large Area Germanium Photocell," *Rev. Sci. Instr.*, **23**, 135 (1952).
78. D. M. Chapin, C. S. Fuller, and G. L. Pearson, "A New Silicon *p-n* Junction Photocell for Converting Solar Radiation into Electrical Power," *J. Appl. Phys.*, **25**, 676 (1954).
79. D. C. Reynolds, G. Leies, L. L. Antes, and R. E. Marburger, "Photovoltaic Effect in Cadmium Sulfide," *Phys. Rev.*, **96**, 533 (1954).
80. M. P. Thekaekara, "Data on Incident Solar Energy," *Suppl. Proc. 20th Annu. Meet. Inst. Environ. Sci.*, 1974, p. 21.
81. C. H. Henry, "Limiting Efficiency of Ideal Single and Multiple Energy Gap Terrestrial Solar Cells," *J. Appl. Phys.*, **51**, 4494 (1980).
82. *Principal Conclusions of the American Physical Society Study Group on Solar Photovoltaic Energy Conversion*, American Physical Society, New York, 1979.
83. M. B. Prince, "Silicon Solar Energy Converters," *J. Appl. Phys.*, **26**, 534 (1955).
84. H. J. Hovel, *Solar Cells*, in R. K. Willardson and A. C. Beer, Eds., *Semiconductors and Semimetals*, Vol. **11**, Academic, New York, 1975; "Photovoltaic Materials and Devices for Terrestrial Applications," *Tech. Dig. IEEE IEDM*, 1979, p. 3.
85. J. Mandelkorn and J. H. Lamneck, Jr., "Simplified Fabrication of Back Surface Electric Field Silicon Cells and Novel Characteristic of Such Cells," *Conf. Rec. 9th IEEE Photovoltaic Spec. Conf.*, IEEE, New York, 1972, p. 66.

86. R. A. Arndt, J. F. Allison, J. G. Haynos, and A. Meulenberg, Jr., "Optical Properties of the COMSAT Non-Reflective Cell," *Conf. Rec. 11th IEEE Photovoltaic Spec. Conf.*, IEEE, New York, 1975, p. 40.

87. H. C. Card and E. S. Yang, "MIS-Schottky Theory under Conditions of Optical Carrier Generation in Solar Cells," *Appl. Phys. Lett.*, **29**, 51 (1976).

88. A. V. Shah, M. Vanecek, J. Meier, F. Meillaud, J. Guillet, D. Fischer, C. Droz, X. Niquille, S. Fay, E. Vallat-Sauvain, V. Terrazzoni-Daudrix, and J. Bailat, "Basic Efficiency Limits, Recent Experiments Results and Novel Light-Trapping Schemes in a-Si:H, μc-Si:H and Micromorph Tandem Solar Cells," *J. Non-Cryst. Solids*, **338-340**, 639 (2004).

89. R. I. Frank, J. L. Goodrich, and R. Kaplow, "A Novel Silicon High-Intensity Photovoltaic Cell," *GOMAC Conference*, Houston, Nov. 1980.

PROBLEMS

1. (a) Show that the quantum efficiency η of a photodetector is related to the responsivity $\mathscr{R}$ at a wavelength λ(μm) by the equation $\mathscr{R} = \eta\lambda/1.24$.

(b) What is the ideal responsivity at a wavelength of 0.8 μm for (1) a GaAs homojunction, (2) an $Al_{0.34}Ga_{0.66}As$ homojunction, (3) a heterojunction formed between GaAs and $Al_{0.34}Ga_{0.66}As$, and (4) a two-terminal, monolithic, series-connected tandem photodetectors when the upper detector is made of $Al_{0.34}Ga_{0.66}As$ and the lower detector is made of GaAs.

2. A photoconductor with dimensions $L = 6$ mm, $W = 2$ mm, and $D = 1$ mm (Fig. 2a) is placed under uniform radiation. The absorption of the light increases the current by 2.83 mA. A voltage of 10 V is applied across the device. As the radiation is suddenly cut off, the current falls, initially at a rate of 23.6 A/s. The electron and hole mobility are 3600 and 1700 cm^2/V-s, respectively. Find (a) the equilibrium density of electron-hole pairs generated under radiation, (b) the minority-carrier lifetime, and (c) the excess density of electrons and holes remaining 1 ms after the radiation is cut off.

3. Calculate the gain and current generated when 1 μW of optical power with $h\nu = 3$ eV is shone onto a photoconductor of $\eta = 0.85$ and a minority-carrier lifetime of 0.6 ns. The material has an electron mobility of 3000 cm^2/V-s, the electric field is 5000 V/cm, and $L = 10$ μm.

4. (a) For a *p-i-n* photodetector, the quantum efficiency is given by Eq. 33. Derive this equation from Eqs. 2 and 32.

(b) A *p-i-n* photodetector has a 1-μm InGaAs absorbing layer. There is an anti-reflection coating (reflectivity = 0%) on the side where the light enters the photodetector.

(1) What is the external quantum efficiency of the photodiode at a wavelength of 1.55 μm?

(2) What would be the external quantum efficiency if light were to travel through the absorbing layer twice?

Assuming the absorption coefficient is 10^4 cm^{-1} at 1.55 μm, and the diffusion length is 10^{-2} cm.

5. For a photodiode, we need a sufficiently wide depletion layer to absorb most the incoming light, but not too wide to limit the frequency response. Find the optimum depletion-layer thickness for Si photodiode having a modulation frequency of 10 GHz.

6. a) For an avalanche photodiode (APD), the breakdown condition is given by Eq. 58. Derive the equation.

b) For a germanium lo-hi-lo APD, if the thickness of the avalanche region is 1 μm, find the electron and hole ionization rates at room temperature.

7. A silicon n^+-p-π-p^+ avalanche photodiode operated at 0.8 μm has a p-layer 3 μm and a π-layer 9 μm thick. The biasing voltage must be high enough to cause avalanche breakdown in the p-region and velocity saturation in the π-region. Find the minimum required biasing voltage and the corresponding doping concentration of the p-region. Estimate the transit time of the device.

8. A p-n junction photodiode can be operated under photovoltaic conditions similar to that of a solar cell. The current-voltage characteristics of a photodiode under illumination are also similar. State three major differences between a photodiode and a solar cell.

9. Consider a silicon p-n junction solar cell of area 2 cm^2. If the dopings of the solar cell are $N_A = 1.7\times10^{16}$ cm^{-3} and $N_D = 5\times10^{19}$ cm^{-3}, and given $\tau_n = 10$ μs, $\tau_p = 0.5$ μs, $D_n =$ 9.3 cm^2/s, $D_p = 2.5$ cm^2/s, and $I_L = 95$ mA, (a) calculate and plot the I-V characteristics of the solar cell, (b) calculate the open-circuit voltage, and (c) determine the maximum output power of the solar cell, at room temperature.

10. At 300 K, an ideal solar cell has a short-circuit current of 3 A and an open-circuit voltage of 0.6 V. Calculate and sketch its power output as a function of operating voltage and find its fill factor from this power output.

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

14

Sensors

14.1 INTRODUCTION
14.2 THERMAL SENSORS
14.3 MECHANICAL SENSORS
14.4 MAGNETIC SENSORS
14.5 CHEMICAL SENSORS

14.1 INTRODUCTION

The human body is equipped with some natural sensors. We are capable of sensing temperature, pressure, light, taste, and so forth. However, sensor devices help greatly to extend our sensitivity as well as the scope of our natural capability to things such as magnetic field. Sensors, explained by its own name, sense or monitor a physical or chemical quantity. Other names have been used for the same or similar purpose, but for the scope of this book, *sensor*, *detector*, and *transducer* are synonymous. Sensors have developed relatively slowly in the semiconductor-device area. But in light of increasing demand for security, environmental control, health improvement, and so on, sensors are expected to become more important and grow at a faster pace in the near future.[1–2]

Figure 1 demonstrates the basic working principle of a sensor. Since this chapter mainly focuses on semiconductor sensors, the output and input are both electrical signals. The measurand is an external influence, property, or condition that is to be detected or measured by a sensor. The measurands can be grouped into the following:

1. Thermal
2. Mechanical
3. Magnetic
4. Chemical
5. Optical

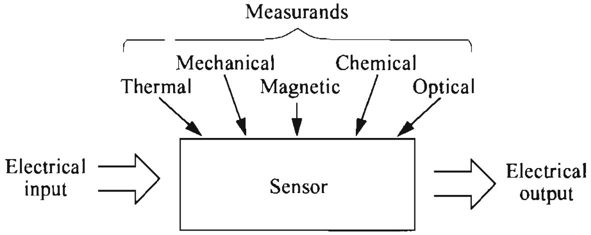


Fig. 1 Application of a common semiconductor sensor is to monitor a physical quantity through the change of electrical signal.

Optical sensors, or photodetectors, have been extensively covered in Chapter 13 and will not be repeated here. Also, if the measurand is another electrical signal, then the sensor becomes a regular semiconductor device and would not be considered in this chapter as a sensor.

In some rare occasions, the sensor does not need an electrical signal input or power. The measurand itself can generate an electrical signal, rather than just modulating it. An example is a photodetector operated under the photovoltaic mode, i.e., without bias a voltage or current can be produced by the light signal. Such sensors are called *self-generating* sensors, as opposed to *modulating* sensors. Most semiconductor sensors, however, are of the modulating type, that is, an input signal or power is required.

It should be pointed out that sensors discussed in the chapter use primary or direct sensing schemes. The actual sensing system in commercial market can be different by using indirect sensing approaches. For example, temperature can be monitored directly by the *I-V* characteristics of a diode, but it can also be monitored by the mechanical expansion of a multi-layered metal whose movement is monitored by a mechanical sensor or optical sensor. The sensing system is not unique and it depends on economics and the environment in use.

As the rest of the book, this chapter focuses mainly on semiconductor material and devices. But for the sake of completeness and the awareness of other choices, important alternatives, if they exist, based on nonsemiconductor materials are briefly mentioned at the end of each sensor group.

14.2 THERMAL SENSORS

14.2.1 Thermistor

The name *thermistor* comes from *thermally sensitive resistor*. There is a long history of the observation of temperature-dependent resistance on different materials, dating back to the nineteenth century. Temperature thermometers using metals, called resistance temperature detectors, will be discussed further in Section 14.2.4. Thermistors usually imply semiconducting materials, and they are of two distinct classes: metal oxides and single-crystal semiconductors. The crystal thermistors are not in competition with those of metal oxides since they cover a different temperature range.

Thermistors can be shaped into different forms, depending on the environment whose temperature is to be monitored. These environments include air ambient, liquid, solid surface, and radiation for two-dimensional imaging. Accordingly, thermistors can be in the form of beads, disks, washers, rods, probes, and thin films. Metal-oxide thermistors are made from fine powders that are compressed and sintered at high temperature. The most-common materials include Mn_2O_3, NiO, Co_2O_3, Cu_2O, Fe_2O_3, TiO_2, and U_2O_3. Single-crystal Ge and Si thermistors are usually doped to 10^{16}–10^{17} cm^{-3}, sometimes with compensating (opposite type) dopants in the order of a few percent.

The range of temperature sensing depends, to the first order, on the energy gap of the materials, that is, larger E_g for higher temperature. Germanium thermistors, which are more common than Si, are used in the cryogenic range 1–100 K. Silicon thermistors are restricted to below 250 K, above which a positive temperature coefficient (PTC) sets in. Metal-oxide thermistors are used in the range 200–700 K. For still higher temperatures, thermistors are made from Al_2O_3, BeO, MgO, ZrO_2, Y_2O_3, and Dy_2O_3.

Since a thermistor is basically a resistor, the conductivity is given by the equation

$$\sigma = \frac{1}{\rho} = q(n\mu_n + p\mu_p). \tag{1}$$

Most thermistors operate in the temperature range in which the ionized concentration (n or p) is a strong function of temperature, given by the form

$$\text{active concentration} \propto \exp\left(\frac{-E_a}{kT}\right) \tag{2}$$

where the activation energy E_a is related to the energy gap and the impurity level. Qualitatively, as temperature goes up, the active doping level goes up and resistance goes down. The decrease of resistance with temperature is called negative temperature coefficient (NTC). Empirically, the net resistance can be described by

$$R = R_o \exp\left[B\left(\frac{1}{T} - \frac{1}{T_o}\right)\right]. \tag{3}$$

R_o is a reference resistance at T_o, and it is common to take room temperature as the reference. B is a characteristic temperature, and it lies in the range 2,000–5,000 K. This factor B has actually a temperature dependence but is weak and can be ignored in a first-order analysis. The temperature coefficient of resistance α is given by

$$\alpha \equiv \frac{1}{R}\frac{dR}{dT} = \frac{-B}{T^2}. \tag{4}$$

The negative sign designates NTC. The change of resistance is the signal arising from a change of temperature ΔT,

$$\Delta R = R\alpha\Delta T. \tag{5}$$

A typical value of α is $\approx -5\%\ \mathrm{K}^{-1}$, which is about 10 times more sensitive than the metal temperature detectors. The resistance of a thermistor typically falls in the range 1 kΩ to 10 MΩ.

At higher temperatures or in heavily doped devices, the dopants are fully ionized, and the decrease of mobility due to phonon scattering starts to dominate the temperature dependence. This gives rise to PTC. Generally, PTC is not as sensitive as NTC and is not utilized in thermistors.

Care has to be taken to avoid self-heating of the thermistor from too high a current. The I-V characteristics resulting from self-heating are different with NTC and PTC. In thermistors with NTC, self-heating induces a drop in resistance and starts a positive feedback for a voltage source (Fig. 2a), leading to higher current. In thermistors with PTC, self-heating increases the resistance and results in negative feedback for a current source (Fig. 2b). These two curves are similar to negative differential resistance with S- and N-shape characteristics.

The advantages of thermistors for temperature measurement include low cost, high resolution, and flexibility in size and shape. The absolute value of resistance is very high, so long cables and contact resistance are more tolerable. The slow response (1 ms to 10 s) is not a critical disadvantage in general applications.

14.2.2 Diode Thermal Sensor

The diode thermal sensor is based on the diffusion current in a p-n junction. Recall the forward-bias diffusion current component in Chapter 2,

$$\begin{aligned} I &= Aq\left(\frac{D_p}{L_pN_D} + \frac{D_n}{L_nN_A}\right)n_i^2\left[\exp\left(\frac{qV}{kT}\right) - 1\right] \\ &\approx Aq\left(\frac{D_p}{L_pN_D} + \frac{D_n}{L_nN_A}\right)n_i^2\exp\left(\frac{qV}{kT}\right) . \end{aligned} \tag{6}$$

Apart from the obvious qV/kT term, both the intrinsic concentration n_i and the ratios D_p/L_p, D_n/L_n are temperature dependent. Since n_i is related to the energy gap, we assume a simplified temperature dependence of

$$E_g(T) = E_g(0) - \alpha T \tag{7}$$

where $E_g(0)$ is the extrapolation to zero temperature (see Section 1.3). From Eq. 28 in Chapter 1, we have

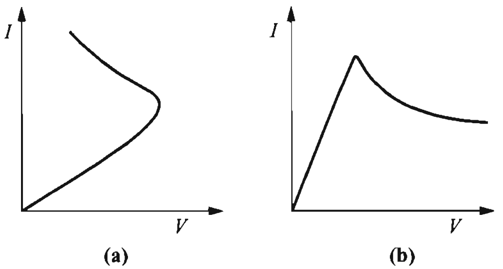


Fig. 2 I-V characteristics of thermistors with self-heating; with (a) negative and (b) positive temperature coefficients.

$$n_i^2 \propto T^3 \exp\left[\frac{-E_g(T)}{kT}\right]$$
$$\propto T^3 \exp\left[\frac{-E_g(0)}{kT}\right]. \tag{8}$$

Next, the diffusion-constant term, depending on the type of carriers (electrons or holes) which dominates the current, has a temperature dependence of (see Section 2.3.1)

$$\frac{D_p}{L_p} \text{ or } \frac{D_n}{L_n} \propto T^{C_1} \tag{9}$$

where C_1 is a constant. Combining these terms into Eq. 6 gives

$$I = C_2 T^{C_3} \exp\left[\frac{qV - E_g(0)}{kT}\right] \tag{10}$$

where C_2 and C_3 are constants. In practice, a known current is passed through the diode, and the terminal voltage is monitored (Fig. 3a). Rearranging Eq. 10 for a voltage expression gives

$$V(T) = \frac{E_g(0)}{q} + \frac{kT}{q}\ln\left(\frac{I}{C_2 T^{C_3}}\right). \tag{11}$$

Since the logarithmic term is insensitive to temperature change, the terminal voltage is linearly dependent on temperature, with an off-set of a constant $E_g(0)/q$. Typical sensitivity is 1–3 mV/°C.

A technique to avoid the constants $E_g(0)$, C_2, and C_3 is to use two bias currents on the same device sequentially or on two identical devices simultaneously. It can be shown from Eq. 11 that the voltage difference between the two measurements is directly proportional to the temperature,

$$\Delta V(T) = \frac{kT}{q}\ln\left(\frac{I_1}{I_2}\right). \tag{12}$$

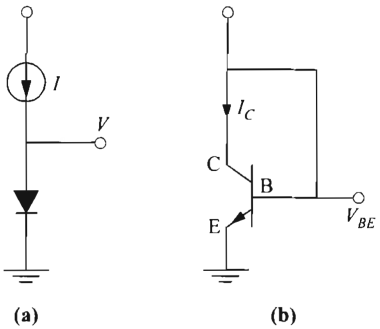


Fig. 3 Measurement of thermal sensor on (a) *p-n* junction diode and (b) bipolar transistor.

14.2.3 Transistor Thermal Sensor

In any *p-n* junction diode, there sometimes exist nonideal currents in addition to the diffusion current. These include surface and bulk recombination currents that add noise to the temperature measurement. Such nonideal effects can be conveniently eliminated by using the collector current of a bipolar transistor. In a bipolar transistor, the emitter current is similar to a *p-n* junction diode current. The collector current, however, filters out the nonideal current components and only consists of the diffusion component. The collector can be shorted to the base as shown in Fig. 3b. By monitoring the collector current, the stringent requirement for an ideal diode is removed, and the fabrication process is more forgiving. The mathematics for a transistor thermal sensor is identical to that of a diode, except now the collector current and the base-emitter voltage V_{BE} are monitored.

14.2.4 Nonsemiconductor Thermal Sensors

Resistance Temperature Detector. The resistance temperature detector (RTD) is similar to a thermistor except it is made of metals. Because of this, it always has positive temperature coefficient and is less sensitive. The most widely used metal is platinum, followed by nickel and copper. The temperature dependence has the general form of

$$R = R_0(1 + C_4T + C_5T^2) \tag{13}$$

where R_0 is the resistance at the reference temperature, usually at 0°C. For platinum, the coefficients are $C_4 = 3.96\times10^{-3}$ /°C and $C_5 = 5.83\times10^{-6}$ /°C. The range of temperature these materials cover are: Pt, –260 to 600°C and up to 900°C with reduced accuracy; Ni, –80 to 300°C; and Cu, –200 to 200°C. The RTD is either in the form of wound wire or foil. It has a resistance of ≈ 100 Ω. Because of the low resistance, four-terminal measurement or a bridge circuit is needed to eliminate parasitic resistance due to connections and contacts.

Thermocouple. The thermocouple is based on thermoelectricity, which is the interaction between thermal energy and electrical energy. There are three thermoelectric effects that are relevant to the operation and fundamental understanding of the thermocouple: the Seebeck effect, the Peltier effect, and the Thomson effect, all named after the respective scientists who made their discoveries in the period 1822–1847. In the Seebeck effect, when two dissimilar wires of conductor or semiconductor are joined together and the two junctions are held at different temperatures, a current arises that flows around the loop (Fig. 4a). When this loop is broken, a voltage can be measured which is sometimes called the Seebeck voltage (Fig. 4b). It can be further demonstrated that this Seebeck voltage can be decomposed into components across each junction and each wire. The Peltier effect states that when a current is passed through a junction, heat is either absorbed or generated, depending on the direction of the current flow. This effect can indeed be used in refrigeration. In the open-circuit condition, a Peltier EMF (V_P) is developed across each junction, and it is a function

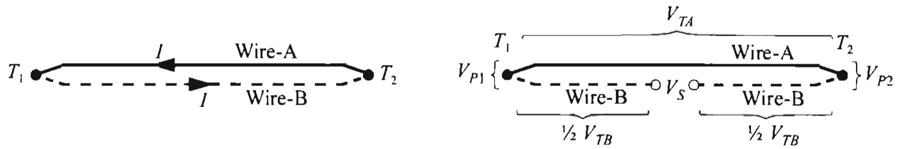


Fig. 4 (a) In a closed thermocouple, when $T_1 \neq T_2$, a circulating current is generated. (b) When the circuit is open, a voltage is developed. This terminal voltage can be decomposed into Peltier EMF (V_P) across the junction and Thomson EMF (V_T) along the wire.

of the temperature. The Thomson effect deals with similar heat exchange from a wire instead of a junction. For an open-circuit condition, when the wire has a temperature gradient along its length, a Thomson EMF is developed. It can be seen from Fig. 4b that the Seebeck voltage V_S is the sum of two Peltier EMFs and two Thomson EMFs, given by

$$V_S = (V_{P1} - V_{P2}) + (V_{TA} - V_{TB}) . \tag{14}$$

The Seebeck voltage is thus a measure of the difference in temperature $(T_2 - T_1)$ of the two junctions. If the junctions are at the same temperature, $V_{TA} = V_{TB} = 0$, $(V_{P1} - V_{P2}) = 0$, and $V_S = 0$.

A thermocouple is used as a temperature sensor. Since the output voltage depends on the difference in the junction temperatures, the temperature of one junction (the reference junction) has to be known. The other junction is then referred to as the sensing or measuring junction. A common reference temperature is 0°C, which can be conveniently obtained by using ice. Room temperature can also be used as reference when accuracy is less critical. The relationship between temperature difference and voltage depends on the thermocouple materials. Lookup tables for this relationship have been established for all thermocouples. The techniques for forming the thermocouple junctions are welding, soldering, and brazing. The choice of thermocouple for an application depends on the suitability of the temperature range. The sensitivity also needs to be considered. It usually falls in the range 5–90 mV/°C.

The thermocouple is widely used as a temperature sensor because it is robust, inexpensive, simple to use, and covers a wide temperature range. The disadvantages are low sensitivity and accuracy, and the need for a reference temperature. The response time of a thermocouple is on the order of ms.

A thermopile simply consists of multiple thermocouples connected in series. The main purpose is to improve the sensitivity since the output voltage is now the sum of all the thermocouple junction pairs.

14.3 MECHANICAL SENSORS

14.3.1 Strain Gauge

The strain of a material is its deformation under stress. A strain gauge (or gage) can measure the strain by monitoring its resistance change. When the strain gauge is elongated, for example, two effects can change its resistance—a geometric effect due to longer length and smaller cross section, and a piezoresistive effect due to a change in resistivity under strain. The latter effect only occurs in semiconductors and is much stronger than the geometric effect. The piezoresistive effect in the semiconductors Si and Ge was discovered by Smith in 1954.[3]

A strain gauge can be made of metal or semiconductor material. The semiconductor strain gauge can be a discrete bonded bar, a diffused or ion-implanted structure, or deposited thin film. The diffused/implanted type is most common because it is compatible with integrated-circuit technology. The semiconductor gauge is usually doped to *p*-type because it has better sensitivity and linearity than *n*-type. It is heavily doped in the 10^{20} cm^{-3} range. Although higher doping decreases the gauge factor (discussed later in more detail), it improves another critical performance—the temperature independence. This trade-off is demonstrated in Fig. 5. Almost all commercial semiconductor strain gauges are made of silicon, although germanium has also been studied. Between semiconductor gauges and metal gauges, the former have much higher sensitivity and higher resistance for reduced power consumption, but the latter have advantages of less temperature dependence, better linearity, higher strain range (4% compared to 0.3%), and better flexibility for attachment to curved surfaces. The most-common metals are copper-nickel alloys such as constantan.

Since the measurement of the strain gauge is resistance, we first derive the relationship between strain and resistance—the piezoresistive effect. Strain S is caused by stress and is the ratio of the change of longitudinal linear dimension to its original length,

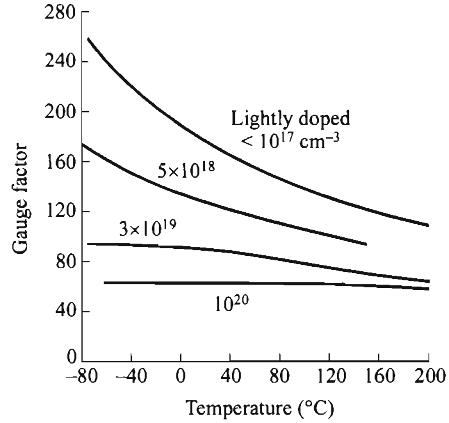


Fig. 5 In semiconductors such as Si, the gauge factor decreases with doping level, but the temperature dependence is lessened. (After Ref. 4.)

$$S = \frac{\Delta l}{l}. \tag{15}$$

The resistance of a bar or film with length l and cross-sectional area A is given by the equation

$$R = \frac{\rho l}{A} \tag{16}$$

When the gauge is under strain, all three parameters—l, A, and resistivity ρ—change, and

$$\begin{aligned}\frac{\Delta R}{R} &= \frac{\Delta l}{l} - \frac{\Delta A}{A} + \frac{\Delta \rho}{\rho} = \frac{\Delta l}{l}\left(1 - \frac{\Delta A/A}{\Delta l/l} + \frac{\Delta \rho/\rho}{\Delta l/l}\right) \\ &\approx S(1 + 2\nu + P_z)\ .\end{aligned} \tag{17}$$

Here ν, Poisson's ratio, relates the longitudinal strain to the transverse strain (linear dimension t is perpendicular to l) by

$$\nu \equiv \frac{-\Delta t/t}{\Delta l/l}. \tag{18}$$

The factor of 2 in Eq. 17 comes from

$$\frac{\Delta A}{A} \approx 2\frac{\Delta t}{t}. \tag{19}$$

P_z is a measure of the piezoresistive effect, which distinguishes a semiconductor gauge from a metal gauge ($P_z \approx 0$). It is given by

$$P_z \equiv \frac{\Delta \rho/\rho}{\Delta l/l} = C_p Y. \tag{20}$$

C_p is the longitudinal piezoresistive coefficient and Y is Young's modulus. The sum

$$G \equiv 1 + 2\nu + P_z = \frac{\Delta R/R}{S} \tag{21}$$

is called the gauge factor. It typically has a value under 2 for metals, but for semiconductors it falls in the range 50–250 and shows improved sensitivity by two orders of magnitude.

In practical operations, strain gauges are incorporated as part of a Wheatstone bridge so that a change in resistance can be detected accurately. The relationship between strain and resistance has to be calibrated. It is usually nonlinear and can be approximated by

$$\frac{\Delta R}{R} = C_6 S + C_7 S^2, \tag{22}$$

where C_6 and C_7 are constants. Also, calibration requires consideration of the temperature dependence of resistivity. It is especially severe for semiconductor gauges. A thermometer mounted near the strain gauge can provide additional data for adjustment. A better approach is to incorporate two or four similar strain gauges in the Wheatstone bridge for automatic temperature compensation, with only one arm exposed to the strain. Other considerations are resistance changes due to self-heating by the measurement bias, and changes due to the photoconductive effect when the gauge is exposed to light.

A strain gauge can be used for many useful mechanical transducers through Hooke's law,

$$S = \frac{T}{Y} \tag{23}$$

where T is the stress. Once the strain is measured, the pressure, force, weight, etc., can be deduced if the Young's modulus of the strained material is known.

Currently, the strain gauge is the most-popular mechanical transducer. Applications of strain gauges can be divided into two types: (1) direct measurements of strain (deformation) and displacement, and (2) indirect measurements of pressure, force, weight, and acceleration through Hooke's law. Major applications are listed below:

1. Direct strain measurements: For structural maintenance such as buildings and bridges, it is often necessary to monitor minute deformations such as bending, stretching, compression, and cracking. Another area of application is in spacecraft and automotive bodies. The monitoring of strain is also mandatory for stress analysis. Measurement of displacement also falls into this category.
2. Pressure transducer: If the strain and stress relationship (Young's modulus) of the gauge material is known, the pressure exerted on the gauge can be measured. A popular pressure transducer for ambient and fluid is the diaphragm type, made of a diffused gauge from silicon, shown in Fig. 6a. The built-in diffused gauge monitors the piezoresistance under differential pressure. The diaphragm is formed by chemical etching of the silicon substrate. This transducer is used in the medical and automotive fields. In a load cell of a weighting scale, the weight is deduced from compression or bending of a shaft onto which a strain gauge is attached or embedded. These load cells are used as heavy-duty truck scales as well as lightweight household electronic scales. The torque of a shaft (torsion bar) can also be measured. Acceleration can be measured via the force (pressure) since

$$\text{force} = \text{mass} \times \text{acceleration}, \tag{24}$$

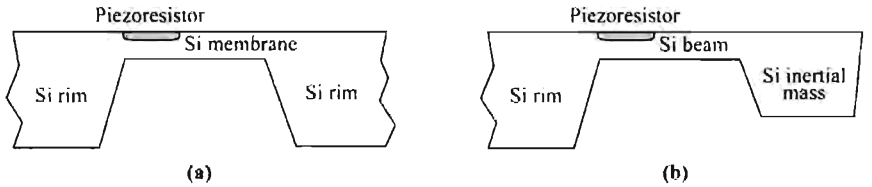


Fig. 6 Sensors based on piezoresistance. (a) Si pressure sensor and (b) acceleration sensor. (After Ref. 1.)

and its implementation is shown in Fig. 6b. Velocity can be deduced from the integration of acceleration. Similar sensors can be made to sense shock, impact, and vibration.

Piezoelectric Strain Gauge. A piezoelectric strain gauge is based on piezoelectricity, an effect that generates a charge and voltage when a piezoelectric crystal is under strain.[5] In operation it is very similar to a piezoresistive strain gauge except that voltage is measured rather than resistance. Structurally, the piezoelectric crystal is sandwiched between two conductive electrodes as shown in Fig. 7. Under stress, the crystal is strained, and a charge or voltage is generated. The process is also reciprocal in that when a voltage is applied, strain and mechanical movement are induced. Good examples for this reciprocal processes are the piezoelectric microphone, where sound pressure produces voltage, and the piezoelectric speaker, where voltage produces strain or mechanical movement (sound wave).

The equations governing piezoelectricity are

$$S = \gamma T + C_{pc}\mathcal{E}, \tag{25}$$

$$\mathcal{D} = C_{pc}T + \varepsilon\mathcal{E}, \tag{26}$$

where γ is the compliance. These state that strain can be created by stress (T) and electric field, and charge (proportional to the electric displacement $\mathcal{D}$) also can be created by the same factors. The piezoelectric charge constant C_{pc} is given by

$$C_{pc} = \frac{Q \text{ per area}}{\text{pressure}}. \tag{27}$$

A piezoelectric transducer is self-generating in that no bias is required, and is dynamic in nature since charge is drained away gradually. For this reason, piezoelectric transducers are more useful in dynamic systems such as accelerometers, loudspeakers, microphones, ultrasonic cleaners, and for sensing shock, vibration, and impact. Other applications include the generation of ignition spark and the positioning of micro-mirrors. (Static application is possible only in the strain-producing mode.) Common piezoelectric materials are quartz, zinc oxide, tourmaline, and ceramics such as lead zirconate titanate and barium titanate. One disadvantage of the

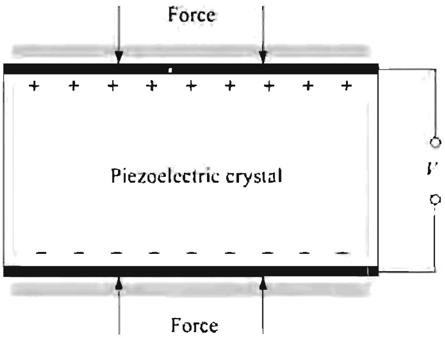


Fig. 7 In a piezoelectric transducer, strain produces charge (and voltage), and vice versa.

piezoelectric transducer is that the source impedance is high, so the first-stage amplifier that senses the voltage must have high input impedance.

14.3.2 Interdigital Transducer

The interdigital transducer (IDT) is a surface-acoustic-wave (SAW) transducer. It converts an electrical signal to a mechanical SAW, and vice versa, based on the piezoelectric effect. For this reason, the IDT is also called an acoustic sensor. The IDT was invented by White and Voltmer in 1965,[6] replacing the older SAW transducers such as the wedge transducer and the comb transducer.

The interdigital transducer consists mainly of interleaved metal fingers on a piezoelectric substrate, as shown in Fig. 8. Alternating fingers are connected to one of two rails. The most critical dimension is the finger period p, which determines the SAW wavelength λ. The linewidth l and space s of the fingers are usually similar and are equal to $p/4$. One common metal is aluminum, with a thickness typically in the 0.1–0.3 μm range; but it should, in any case, be less than $l/2$. The overlap of the metal fingers W can vary even within one IDT. This is especially common in the output IDTs for the purpose of signal processing, and such a structural approach is called *apodisation.* The number of finger pairs N depends on the application. Large N produces more efficient coupling between the electrical signal and the SAW, but the bandwidth also suffers, as discussed later. Common piezoelectric materials are quartz, $LiNbO_3$, ZnO, $BaTiO_3$, $LiTaO_3$, and lead zirconate titanates. These materials are also good insulators. Less common piezoelectric materials are semiconductors such as CdS, CdSe, CdTe, and GaAs. A prerequisite of the piezoelectric effect is some degree of lattice order, so crystal or polycrystal structures are required. For thin-film IDTs, the piezoelectric film thickness is on the order of the SAW wavelength, and ZnO is the most-common material, deposited by sputtering. The piezoelectric thin film can be either under or over the metal layer.

In most applications, two IDTs are used, one in converting an electrical input signal to a SAW that propagates through a medium, and another in converting the SAW back to an electrical signal (Fig. 9). These packages of two IDTs plus the medium are called SAW devices. By monitoring the characteristics of the transmitted

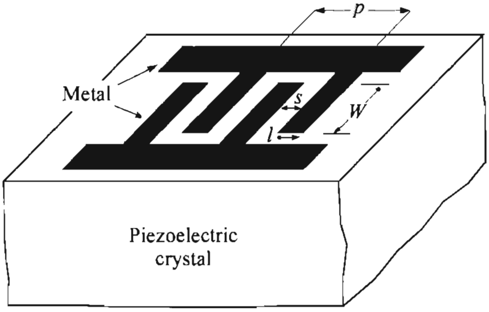


Fig. 8 Interdigital transducer on bulk piezoelectric substrate. Deposited piezoelectric thin film under or over the metal IDT is also possible.

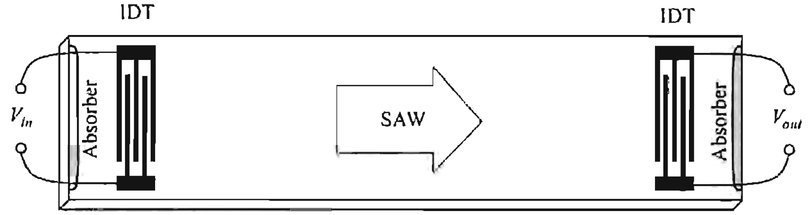
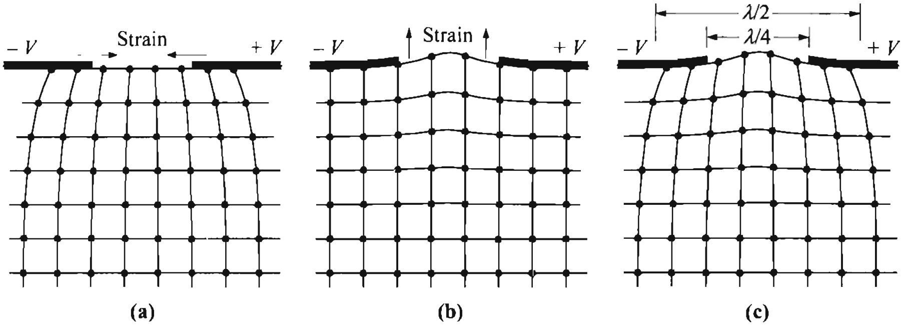


Fig. 9 In acoustic sensing, an IDT generates SAW that passes through the medium which affects the SAW, and another IDT converts SAW back to electrical signal. The absorbers at both ends minimize reflected waves.

SAW, some properties of the medium can be known. The success of IDT led to the dominance of SAW devices over the bulk-acoustic-wave (BAW) devices.

The main function of an interdigital transducer is to interchange energy between an electrical signal and a SAW. To help visualize a SAW, one good analogy is the propagating ripple generated by throwing a stone into calm water, or by a moving boat. In a solid, a SAW is due to the deformation of the structure, or strain. Microscopically, atoms in a crystal are displaced from their equilibrium positions, and the restoring force, similar to that of a spring, is proportional to their displacement. For this reason, a SAW is also called an elastic wave. A SAW differs from a BAW in that it travels along the surface, with most of the energy confined within a wavelength of the surface. A SAW can be separated into a longitudinal wave, where the atom displacement is parallel to the direction of wave propagation, and a shear wave, where this displacement is perpendicular to the wave propagation (Fig. 10). Whether the SAW generated is predominately a longitudinal wave or a shear wave depends on the piezoelectric properties and the crystal orientation. Figure 11 shows the origin of the piezoelectric effect and the fact that the relationship between polarization of the charge and strain depends on the crystal structure.



(a) (b) (c)

Fig. 10 Representation of SAW (parallel to the surface) by displacement of atoms under the influence of an interdigital transducer. (a) Longitudinal wave. (b) Shear wave. (c) Composite of the two waves.

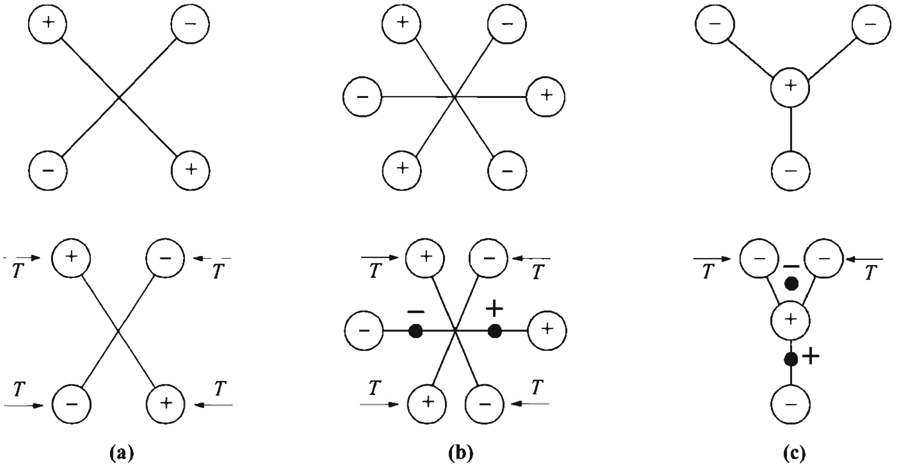


Fig. 11 Origin of piezoelectric effect, showing polarization due to stress T. (a) Stress produces no polarization in a symmetric crystal. (b) Polarization parallel to stress. (c) Polarization perpendicular to stress. (After Ref. 5.)

The velocity of the SAW propagation v depends on the elastic stiffness and the mass density of the medium. For all the practical piezoelectric materials mentioned above, it falls in the range 1–10×10^5 cm/s, with most around 3×10^5 cm/s. The frequency response of an IDT centers at a frequency of

$$f_o = \frac{v}{\lambda} = \frac{v}{p}. \tag{28}$$

A small finger period p can therefore accommodate high-frequency operation, in spite of the low velocity. The frequency response of an IDT is given by

$$R(f) = C_8\frac{\sin X}{X} \tag{29}$$

where C_8 is a constant and

$$X = N\pi\left(\frac{f-f_o}{f_o}\right). \tag{30}$$

This frequency response is shown in Fig. 12. The dependence of bandwidth on the finger-pair number is evident here.

The attractiveness of a SAW device is its low characteristic velocity, five orders of magnitude slower than an electromagnetic wave. Very large delay can be obtained with a reasonable size. A typical delay is ≈ 3 ms/cm. Slow velocity also translates into small wavelength and physical dimension (Eq. 28). It is rather interesting to note that a microwave circuit of ≈ 5 GHz requires that the lateral transistor dimension be scaled to ≈ 0.3 μm, and such frequency of operation for an IDT also requires similar linewidth and space dimension. Other advantages of SAW devices include low atten-

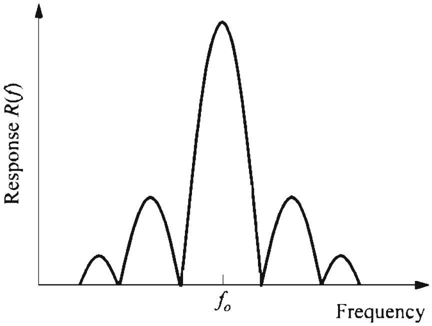


Fig. 12 Frequency response of interdigital transducer without apodisation.

uation, low dispersion (velocity variation with frequency), easy access of SAW, and compatibility with integrated-circuit technology. BAW devices lose some of these advantages. They are used only for frequencies below 10 MHz, which require unreasonably large SAW devices.

The applications of SAW devices lie in two main areas: sensing and signal processing. For sensing applications, the delay or magnitude modulation of the SAW is critical and informative. The velocity and magnitude of the SAW in the sensing area are affected by the physical quantities of the sensing medium such as temperature, moisture, pressure, and stress (Fig. 9). Gas flow can be detected by sensing the cooling effect. Also, if the sensing area is coated with a special absorbant, the SAW is sensitive to certain chemicals and gases such as H_2, SO_2, NO_2, and NH_3.[7] When two IDTs are deposited on a surface, nondestructive testing on this surface for cracks and other defects can be performed. Finally, since SAW is a mechanical wave, light can be diffracted from the surface like a grating. This property is used in diagnostics, optical modulators, and light deflectors.

For signal processing, the most-common applications are delay lines and bandpass filters. Absorbers at the ends are needed because the SAW generated by an IDT is bidirectional (Fig. 9). The absorber also costs a 3-dB loss of transmitted power. Other functional SAW devices include pulse-compressor (chirp filter), oscillator, resonator, convolver, correlator, etc. These SAW devices are useful in communication, radar, and broadcasting equipment such as TV receivers.

14.3.3 Nonsemiconductor Mechanical Sensor

Capacitive Sensor. A simple mechanical structure to monitor pressure is a capacitive sensor which detects the spacing between the two conducting plates (Fig. 13). The capacitance has a linear relationship with the distance d

$$C = \frac{A\varepsilon}{d} \tag{31}$$

where ε is the permittivity in the space between the electrodes. However, the relationship between pressure and displacement has to be known. This depends on the material properties of the cantilever that holds the movable electrode. In this respect, the

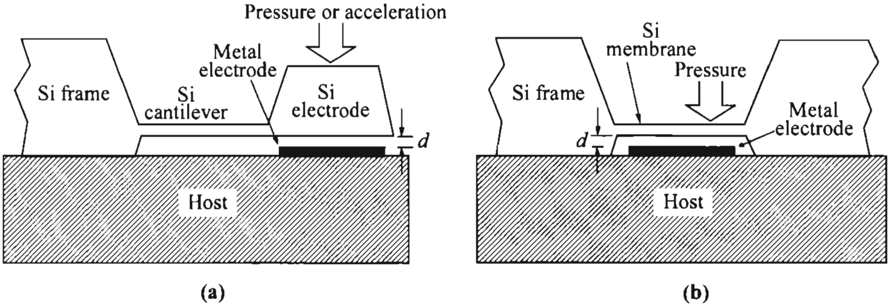


Fig. 13 In a capacitive sensor, the distance between the electrodes is measured by their capacitance. (a) Pressure or acceleration sensor using Si cantilever. (b) Pressure sensor using Si membrane.

materials do not have to be made of semiconductors. However, commonly silicon is used as the cantilever, taking advantages of the mature and inexpensive silicon processing technologies. Another configuration can be made of a semiconductor diaphragm which has been thinned down from the wafer thickness and serves as the flexible electrode.

14.4 MAGNETIC SENSORS

The main applications of magnetic sensors can be grouped into two functions; direct magnetic-field sensing, and position and motion sensing. Equipment to measure the magnet-field strength is called a magnetometer or gaussmeter. Special applications are pickup heads for magnetic tapes (including the magnetic strips in credit cards), magnetic disks, and bubble memories. Also, since a dc or ac current produces a magnetic field in the vicinity of the wire, the current can be detected indirectly. This is advantageous compared to the use of a regular ammeter which has to be inserted in series with the wire. In the second group of applications, when a magnet is attached to an object, its position, displacement, and angular sensing are possible. Examples of applications in angular sensing are the tachometer, dc brushless motor, and for timing automobile engines for spark plugs. A contactless switch can be made from proximity sensing when a magnet is moved in and out of a magnetic sensor. Examples are switches for computer keyboards and closed-loop security systems.

Even though there are many variations of magnetic sensors, they are all based on the Hall effect and readers can refer to Section 1.5.2 for a detailed review.

14.4.1 Hall Plate

The Hall plate is also called the Hall generator. Since the Hall effect is very weak in metals, it was not practical until the realization of good semiconductor materials.

Commercial Hall plates were available as discrete sensors in the mid-1950s and as integrated sensors around 1970.

The Hall plate is simply a piece of semiconductor with four contacts. It exists in one of the forms: (1) discrete bar, (2) thin film deposited on a supporting substrate, and (3) doped layer on an opposite-type substrate. The schematic structure of a Hall plate from integration-circuit technology is shown in Fig. 14. The doping of the active layer should be minimized to maximize the Hall voltage V_H which is inversely proportional to the concentration. Common materials used are InSb, InAs, GaAs, Si, and Ge. Compound semiconductors are attractive for their high mobilities, while Si is more popular for integrated sensors because of its more-mature technology.

The Hall effect is the generation of a Hall voltage V_H when a piece of semiconductor is biased with a current and placed under a magnetic field that is orthogonal to the current flow. The generated Hall voltage, assuming a Hall factor $r_H = 1$ and a p-type semiconductor, is given by

$$V_H = R_H W J_x \mathscr{B} = W \mathscr{E}_x \mu \mathscr{B} \ . \tag{32}$$

Note that to obtain a large signal, the carrier concentration must be minimized for a large R_H. This is the reason that the Hall effect is much more pronounced in semiconductors than in metals.

The sensitivity of a Hall plate has various definitions, depending on whether it is current-related, voltage-related, or power-related. These are given by $\partial V_H / I \partial \mathscr{B}$, $\partial V_H / V \partial \mathscr{B}$, $\partial V_H / P \partial \mathscr{B}$, or simply $\partial V_H / \partial \mathscr{B}$. In any case, an efficient Hall plate should have low carrier concentration and high carrier mobility. Typical sensitivity is ≈ 200 V/A-T, but values up to 1,000 V/A-T are possible. Materials of higher mobilities such as GaAs and InP are preferable over Si in this respect. More-recent structures use heterojunctions, modulation-doped channels, and quantum wells for superior mobilities.

The length L should have a minimum value of $3 \times W$ so that the geometry effect does not diminish the Hall voltage substantially. Physically, it means that if L is too short compared to W, carriers reach the opposite terminal without a sufficient chance

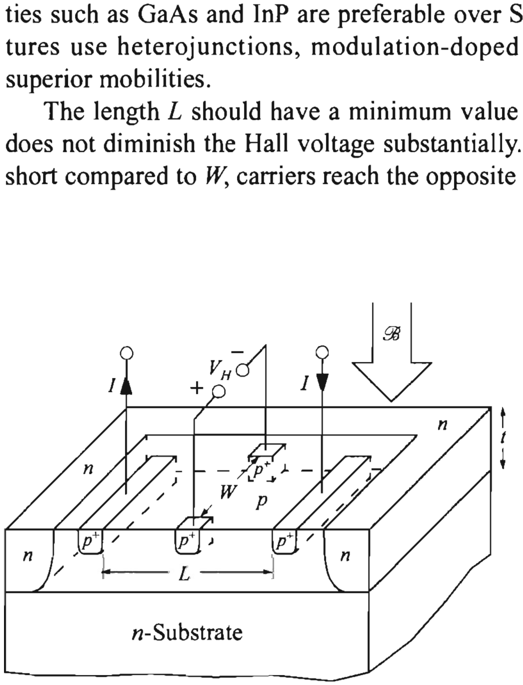


Fig. 14 Structure of a Hall plate from integrated-circuit technology. Surrounding n-regions form isolation.

to be deflected to the sides to develop the full Hall voltage. This effect is accounted for by the geometric correction factor ($G < 1$) such that

$$V_H = GR_H W J_x \mathscr{B} = GW\mathscr{E}_x \mu \mathscr{B} . \tag{33}$$

G is a function of the L/W ratio and is plotted in Fig. 15.

As a magnetic-field sensor, it is critical that V_H be proportional to $\mathscr{B}$ linearly and with zero offset, i.e., V_H goes to zero when $\mathscr{B} = 0$. In practice, there is often an offset voltage when $\mathscr{B} = 0$. The source of this offset is from both a geometric effect and a piezoelectric effect. The geometric effect is due to the fact that the two Hall taps are not exactly opposite to each other. If there is a misalignment Δx between them in the direction of the current flow, the offset voltage is given by

$$\Delta V_H = \mathscr{E}_x \Delta x \; . \tag{34}$$

The piezoelectric effect is the generation of a voltage when a piezoelectric material is under stress. This is especially severe for thin-film Hall plates. Offset can also be due to piezoresistivity and temperature variation. The offset voltage can be eliminated by connecting two or four Hall plates together in a configuration that cancels the individual offset voltages, or by adding a fifth terminal as a control gate to inject current for compensation.

The Hall plate is attractive for its low cost, simple structure, and compatibility with integrated-circuit technologies.

14.4.2 Magnetoresistor

The magnetoresistor is based on the magnetoresistance effect, which is an increase of resistance in the presence of a magnetic field. The magnetoresistance effect arises from two independent mechanisms: (1) a physical magnetoresistance effect and (2) a geometric magnetoresistance effect.

The physical magnetoresistance effect arises because carriers do not move with identical velocity. The Hall voltage is set up to balance an average velocity, and carriers with velocities that differ from the average would deviate from the shortest path, as shown in Fig. 16a. These longer paths lead to increased resistance. This physical magnetoresistive effect results in the general magnetic-field dependence of

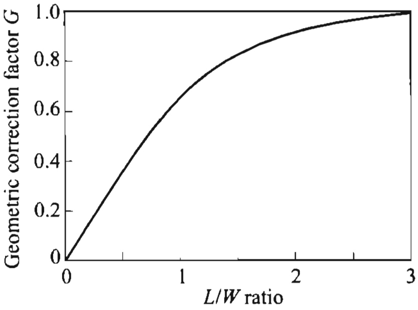


Fig. 15 Geometric correction factor as a function of the L/W ratio. (After Ref. 2)

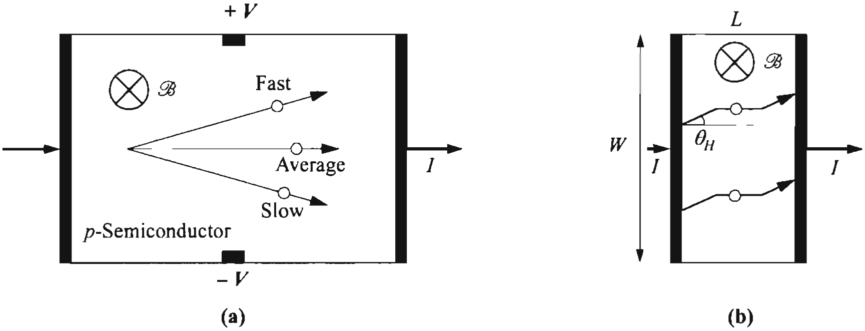


Fig. 16 (a) Physical magnetoresistance effect is caused by carriers having nonuniform velocities. Carriers with higher and lower speed than the average have to travel longer distance. (b) Geometric magnetoresistance effect occurs in samples with small L/W ratios. Carriers near the contacts move at the Hall angle.

$$R(\mathscr{B}) = R(0)(1 + C_9\mu^2\mathscr{B}^2) \tag{35}$$

where C_9 is a constant.

The geometric magnetoresistance effect occurs in samples with small L/W ratios. In this case, the full Hall voltage is not fully developed to balance the Lorentz force (Eq. 33 and Fig. 16b), and carriers near the contacts move at an angle to the applied electric field. The longer path again leads to higher resistance. A magnetoresistor maximizing this effect is shown in Fig. 17a, where conductive shorts are added and the structure is equivalent to many Hall plates in series, each having a small L/W ratio. Another magnetoresistor is in the form of a Corbino disc. The contacts are concentric such that there is no sides for the Hall voltage to develop. To calculate the angle of

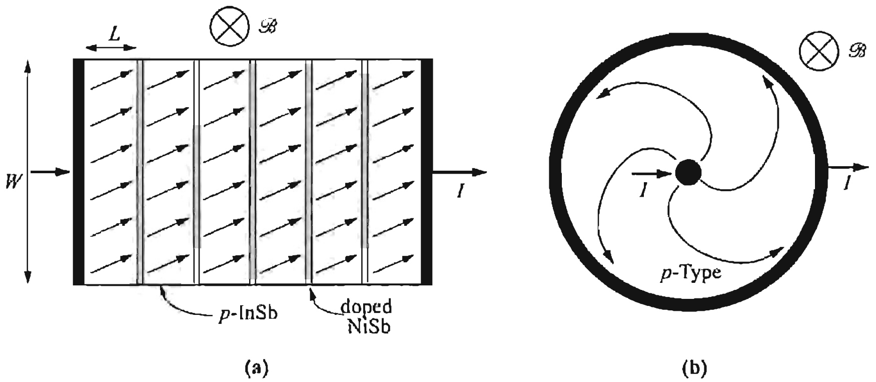


Fig. 17 Magnetoresistors maximizing the geometric magnetoresistance effect. (a) Highly doped shorts divide the sample into regions of small L/W ratio. (b) The Corbino disc does not allow a Hall voltage to develop. Arrows show paths of holes.

movement, since in a regular Hall plate, carriers move with a straight path in the presence of the Hall field $\mathscr{E}_H$, it is reasonable to assume that without this Hall field, the carrier path would make an angle called the Hall angle,

$$\theta_H \equiv \tan^{-1}\left(\frac{\mathscr{E}_H}{\mathscr{E}_x}\right) = \tan^{-1}(r_H\mu\mathscr{B}). \tag{36}$$

The resultant resistance would become

$$\begin{aligned} R(\mathscr{B}) &= R(0)(1 + a\tan^2\theta_H) \\ &= R(0)(1 + ar_H^2\mu^2\mathscr{B}^2) \qquad 0 < a < 1 . \end{aligned} \tag{37}$$

The factor a is added for the extent of this geometric effect. It is unity in the extreme of small L/W ratio (< 1/4), and vanishes when the ratio is large (> 4). The square term comes about because the current path not only is longer, but is also narrower.

14.4.3 Magnetodiode

A magnetodiode is a *p-i-n* diode that contains a surface of high recombination rate in the intrinsic layer (Fig. 18). When the *p-i-n* diode is under forward bias, the intrinsic layer has high concentrations of injected electrons and holes, and the current is controlled by recombination. Under a magnetic field, both electrons and holes are deflected toward the same surface of high recombination rate, and the recombination current is increased. The purpose of the middle intrinsic layer is to have a large depletion layer to maximize this recombination effect. A practical magnetodiode can be made from an SOS (silicon-on-sapphire) film where the bottom Si/Al_2O_3 interface naturally has a higher density of defects, for the detection of a magnetic field parallel to the surface. The disadvantages associated with the magnetodiode is poor reproducibility, poor linearity, and poor temperature dependence.

14.4.4 Magnetotransistor

A magnetotransistor, also called a magnistor, usually implies a bipolar transistor with multiple collector contacts whose current difference is proportional to the magnetic field. These bipolar transistors can be lateral or vertical structures. Each can also

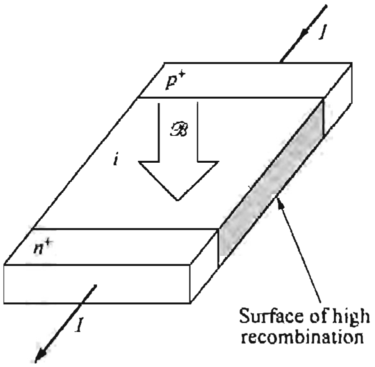


Fig. 18 A magnetodiode is a *p-i-n* diode having one surface with high defect density.

operate in the deflection mode or injection-modulation mode. This combination of four schemes is represented by the top views and cross sections in Fig. 19. In the deflection mode of operation, the injected carriers are deflected in the base or collector region by the magnetic field, and they get collected unevenly by the two collector contacts. In the injection-modulation mode, the base has two contacts and acts as a Hall plate. Under a magnetic field, the base has unequal potential locally and causes uneven emitter-base bias and thus uneven emitter injection. This also leads to uneven collector currents. The difference in collector currents is proportional to the magnetic field, given by

$$\Delta I_C = K\mu I_C \mathscr{B} . \tag{38}$$

The factor K depends on the geometry and bias condition of the device.

14.4.5 Magnetic-Field-Sensitive Field-Effect Transistor

The magnetic-field-sensitive field-effect transistor (MAGFET) usually implies a MOSFET structure. It can operate in two modes with different structures. The structure in Fig. 20a is similar to a Hall plate where the sample thickness t is replaced by a surface induced inversion layer, and the output is the Hall voltage across the Hall taps. This Hall voltage is given by an expression similar to the Hall plate,

$$V_H = \frac{G r_H I_D \mathscr{B}}{Q_{in}} = \frac{G r_H I_D \mathscr{B}}{C_{ox}(V_G - V_T)} . \tag{39}$$

Here the parameters represent those of a regular MOSFET, where Q_{in} is the inversion-layer charge sheet, and V_T is the threshold voltage. The device operates in the linear regime $[V_D \ll (V_G - V_T)]$. Compared to a regular Hall plate, the MAGFET suffers from lower surface mobility and higher $1/f$ noise. It has the advantage of variable carrier concentration.

The split-drain MAGFET is shown in Fig. 20b. Under a transverse magnetic field, carriers in the MOSFET channel are deflected toward one side, and the difference between the two drain currents is monitored. The operation is similar to that of the lateral magnetotransistor shown in Fig. 19a, and the characteristics can be described by that similar to Eq. 38.

14.4.6 Carrier-Domain Magnetic-Field Sensor

A carrier domain is a plasma of electrons and holes. It can be created, for example, by turning on a *p-n-p-n* structure, as in a thyristor. A vertical carrier-domain magnetic-field sensor is shown in Fig. 21a. Due to the symmetry of the device, the carrier domain is formed at the center. Under a magnetic field, the carrier domain is shifted laterally, imposing a change of currents between I_{p1} and I_{p2} and between I_{n1} and I_{n2}. The disadvantages of this sensor is its temperature dependence. Another version of horizontal and circular carrier-domain magnetometer has been studied (Fig. 21b). In this arrangement, the carrier domain rotates around a circle with a frequency that is proportional to the magnetic field. The detection of the domain is through the outermost segmented collectors. This output of frequency is a unique feature of the sensor.

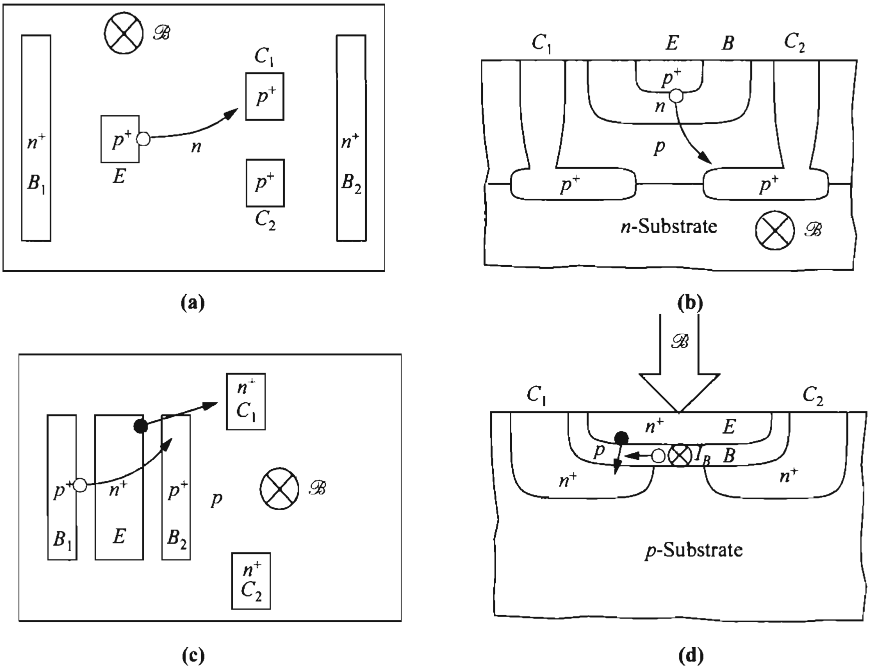

ℬ
C_1
p^+
n^+
B_1
p^+
n
E
p^+
C_2
n^+
B_2
(a)
C_1
E
B
C_2
p^+
n
p
p^+
p^+
n-Substrate
ℬ
(b)
n^+
C_1
p^+
n^+
p^+
p
ℬ
B_1
E
B_2
n^+
C_2
(c)
ℬ
C_1
C_2
n^+
E
p
I_B
B
n^+
n^+
p-Substrate
(d)

Fig. 19 (a) (c) Top views and (b) (d) cross sections of magnetotransistors. (a) Lateral magnetotransistor with deflection mode. Two terminals for the base are for driving the carriers in the base to higher velocity. (b) Vertical magnetotransistor with deflection mode. (c) Lateral magnetotransistor with injection-modulation mode. (d) Vertical magnetotransistor with injection-modulation mode. E = emitter, B = base, C = collector.

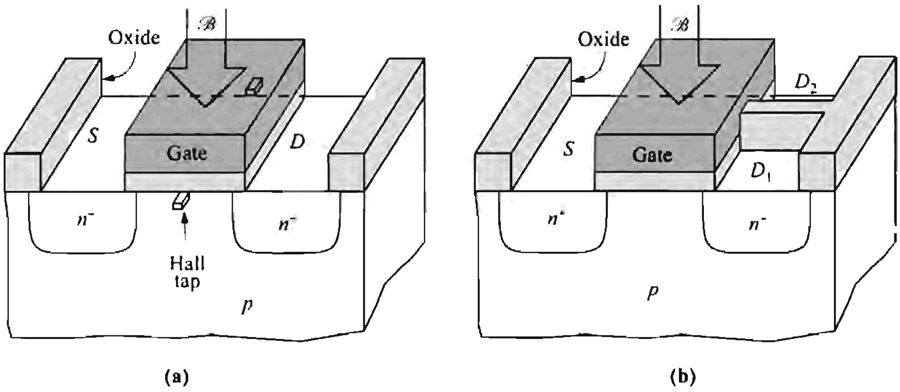

Oxide
ℬ
S
Gate
D
n^-
n^-
Hall
tap
p
(a)
Oxide
ℬ
D_2
S
Gate
D_1
n^-
n^-
p
(b)

Fig. 20 (a) MAGFET using the inversion channel as the Hall plate. (b) Split-drain MAGFET.

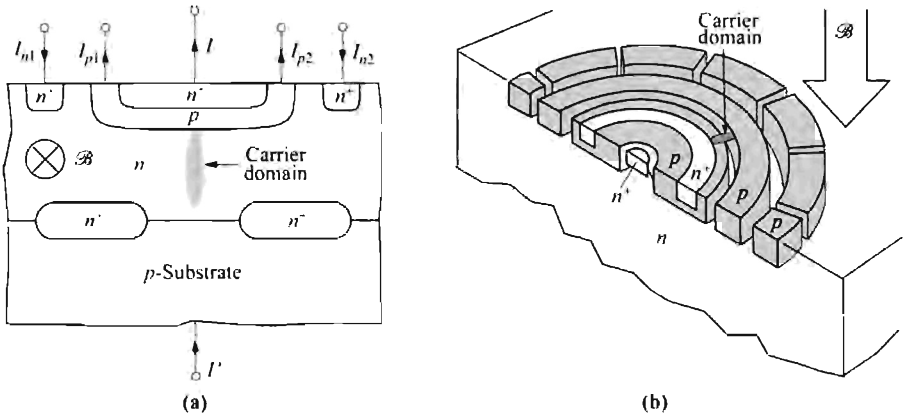


Fig. 21 (a) Carrier-domain magnetometer. Under a magnetic field, the carrier domain shifts laterally, and the current differences $I_{p1} - I_{p2}$ and $I_{n1} - I_{n2}$ are detected. (b) Horizontal, circular carrier-domain magnetometer. The carrier domain rotates with a frequency proportional to magnetic field. (After Ref. 8.)

14.5 CHEMICAL SENSORS

14.5.1 Metal-Oxide Sensors

Gas sensors can be made from metal-oxide semiconductors such as SnO_2, Fe_2O_3, TiO_2, ZnO, In_2O_3, and WO_3, with SnO_2 being the most common.[9–10] These resistive gas sensors are in polycrystal form and made from either powders sintered at high temperature or deposition on some substrate by evaporation or sputtering. Often, some noble metal such as Pd or Pt is added to improve their sensitivity. When exposed to a specific gas the resistance changes. The gas species that can be monitored include H_2, CH_4 (methane), O_2, O_3 (ozone), CO, CO_2, NO, NO_2, SO_2, SO_3, HCl, etc. The sensitivity of these semiconducting-oxide sensors can usually be improved by operation above room temperature, in the range 200–400°C. One reason being that their energy gaps are high, in the range of 3–4 eV, so their resistance can be reduced at higher temperatures to a more practical value for the measurement.

The mechanism responsible for the change of resistance is believed to be due to reactions at the grain boundaries. In many cases, the sensitivity can be shown to improve with smaller grain size for higher gain-boundary density. A few models have been proposed for the resistance change upon exposure to gas elements. Here we mention the two most-popular theories. The first is related to conduction across gain boundaries. These grain boundaries are oxygen rich, and potential barriers are formed that deplete the carriers surrounding them and impede the current flow across them (Fig. 22). The gas to be detected can neutralize the preabsorbed oxygen, reduce the barrier, and reduce the resistance.

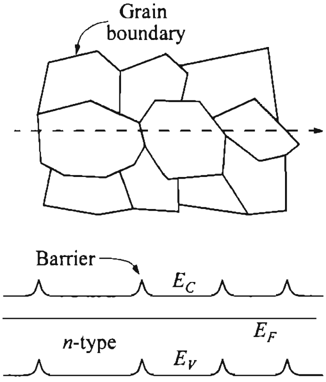


Fig. 22 In a metal-oxide semiconductor sensor, the grain boundaries give rise to potential barriers.

Another possible mechanism is a bulk effect. Here the gas molecule reacts with the absorbed oxygen at the grain boundary, and a free electron is released or neutralized, depending on the type of reaction. This process changes the net carrier concentration in the bulk and its resistance.

In spite of some problems in nonideal reproducibility, long-term stability, sensitivity, and selectivity, these metal-oxide gas sensors have a substantial commercial market because they are very inexpensive and simple to use.

14.5.2 Ion-Sensitive Field-Effect Transistor

The ion-sensitive field-effect transistor (ISFET) is one of the most common of the chemically sensitive field-effect transistors. The ISFET was proposed and demonstrated by Bergveld in 1970.[11–12] Since the inclusion of a reference electrode in contact with the electrolyte was reported in 1974,[13] such an electrode has been considered to be an integral part of an ISFET. Since the function of the ISFET is to detect ions, an electrolyte containing the ions has to be in contact with the transistor. In this arrangement, the electrolyte becomes the gate of a MOSFET (Fig. 23), replacing the conventional poly-Si gate. The contact to the electrolyte gate is provided by a reference electrode, typically Ag-AgCl. The gate dielectric is a critical part of the struc-

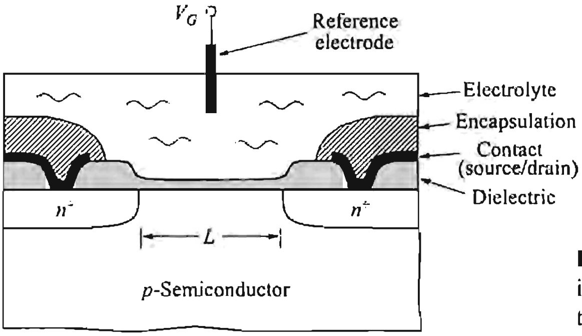


Fig. 23 ISFET (n-channel) immersed in an electrolyte containing the measurand.

ture, and often a multilayered gate dielectric is used. A barrier layer on top of SiO_2 is sometimes necessary to prevent ions from penetrating to the SiO_2/Si interface. The top layer of dielectric is chosen to maximize the sensitivity as well as the selectivity of the ions of interest. Examples of these dielectrics are Si_3N_4, Al_2O_3, TiO_2, and Ta_2O_5. A major concern in the design of an ISFET is the encapsulation, which should prevent ions from penetrating to the rest of the device structure. Typical dimensions of the channel length L and width W are tens to hundreds of microns.

To understand the operation of the ISFET, it is best to start with a conventional MOSFET (see Chapter 6). The electrical characteristics can be roughly divided into two regimes—linear and saturation—and their I-V characteristics are described by

$$I_{lin} = \frac{\mu C_i W(V_G - V_T)V_D}{L}, \tag{40}$$

$$I_{sat} = \frac{\mu C_i W}{2L}(V_G - V_T)^2. \tag{41}$$

The criterion separating these regimes is given by the drain bias

$$V_{D,sat} = V_G - V_T. \tag{42}$$

An important parameter for any FET is the threshold voltage V_T. It is the gate voltage required to turn the transistor on, and is given by

$$V_T = V_{FB} + 2\psi_B + \frac{\sqrt{2\varepsilon_s qN(2\psi_B)}}{C_i} \tag{43}$$

where

$$V_{FB} = \phi_m - \phi_s \tag{44}$$

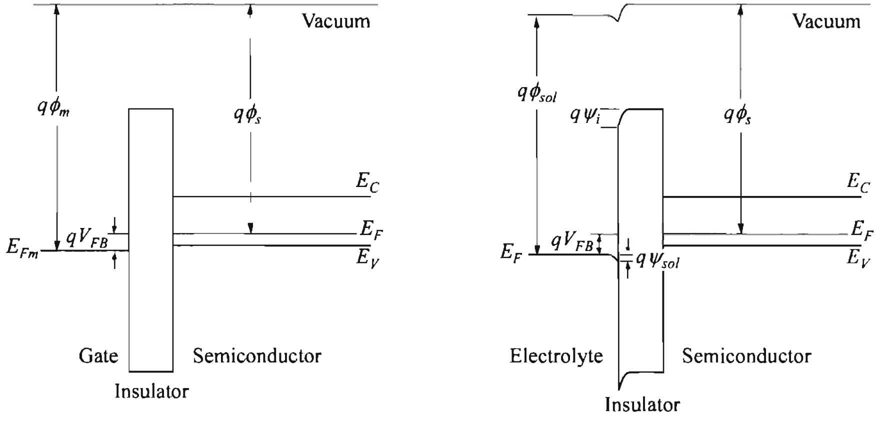
is the flat-band voltage. A metal work function ϕ_m is used since a metal gate is assumed.

These equations are applicable to an ISFET, with the exception of Eq. 44. The difference is explained with the energy-band diagrams in Fig. 24 at flat-band conditions. For an ISFET, it can be seen that

$$V_{FB} = \phi_{sol} - \phi_s + \psi_i - \psi_{sol} \tag{45}$$

where ψ_i is the surface potential of the insulator due to a dipole layer at the dielectric side of the electrolyte/dielectric interface, and ψ_{sol} is the potential drop at the solution side of the same interface. Furthermore, ψ_{sol} is insensitive to ions. The detection of ions relies on the change of ψ_i with ion concentration. In effect, ions get deposited on the insulator surface and change ψ_i, V_{FB}, V_T, and the FET current. The presence of ions is equivalent to a change of gate bias. In practice, the ISFET is biased with a constant source-to-drain current I_D, and the change of gate voltage to sustain such current is the indicator. Examples of ions that are detectable are H^+ (pH), Na^+, K^+, Ca^{2+}, Cl^-, F^-, NO^{3-}, and CO_3^{2-}. Typical values are 20–40 mV/pH for pH sensing.

Compared to other electrochemical ion sensors, the ISFET has the advantages of small size, fast response, low output impedance, and low cost due to integrated-circuit technology. It is now commercially available. The main applications currently



(a) (b)

Fig. 24 Energy-band diagrams at flat-band conditions for (a) a conventional MOSFET, and (b) an ISFET in contact with an electrolyte.

are in the biomedical field. For example, in blood and urine analysis, components such as pH, Na^+, K^+, Ca^{2+}, Cl^-, glucose, urea, and cholesterol can be monitored. The limitations of long-term reliability and irreversibility are important concerns. Because of the nature of these applications, most ISFET sensors are disposable.

14.5.3 Catalytic-Metal Sensors

There is a group of sensors that utilize the change of work function of catalytically active metals when exposed to certain gases.[14] These catalytic-metal sensors are part of the semiconductor devices in the forms of: (1) MOSFET, (2) MOS capacitor, (3) MIS tunnel diode, and (4) Schottky-barrier diode. For a MOSFET, the catalytic metal is used as the gate material. The change in its work function changes the threshold voltage and thus the MOSFET current. For the MOS capacitor, since the capacitance varies with the gate bias, a change in work function causes a shift in the *C-V* curve. For the other two devices, the MIS tunnel diode and the Schottky-barrier diode, the barrier heights are modified and the forward currents are affected accordingly.

The catalytic metals can be Pd, Pt, Ir, and Ni, with Pd being by far the most successful. The most effective detection using this property is on hydrogen gas. The mechanism is believed to be due to the adsorption of H_2 gas by the catalytic metal. H_2 molecules then dissociate into H^+ ions and diffuse to the metal interface, which is in contact with the rest of the device, and a dipole layer is formed. This dipole layer changes the effective work function of the metal.

14.5.4 Biosensors

Biosensors are considered part of the chemical sensors. In fact, a biosensor is the integration of a biologically active membrane in contact with a conventional sensor. The type of sensor used depends on the type of biological reaction monitored. If the product or byproduct is to be detected, the sensor would be one of the chemical sensors discussed above. Otherwise the sensor can be a thermal sensor if there is a heat exchange in the reaction, or it can be a photodetector if the optical absorption is changed, and so on. Because the device part is the same and the biological reaction is not relevant to device engineering, we would not cover the biosensors and readers are referred to literature for further study.[15]

REFERENCES

1. S. M. Sze, *Semiconductor Sensors*, Wiley, New York, 1994.
2. S. Middelhoek and S. A. Audet, *Silicon Sensors*, Academic Press, London, 1989.
3. C. S. Smith, "Piezoresistance Effect in Germanium and Silicon," *Phys. Rev.*, **94**, 42 (1954).
4. W. P. Mason, "Use of Solid-State Transducers in Mechanics and Acoustics," *J. Audio Eng. Soc.*, **17**, 506 (1969).
5. A. J. Pointon, "Piezoelectric Devices," *IEE Proc.*, **129**, Pt. A, 285 (1982).
6. R. M. White and F. W. Voltmer, "Direct Piezoelectric Coupling to Surface Elastic Waves," *Appl. Phys. Lett.*, **7**, 314 (1965).
7. J. W. Grate, S. J. Martin, and R. M. White, "Acoustic Wave Microsensors," *Anal. Chem.*, **65**, Part I, 940A, Part II, 987A (1993).
8. H. P. Baltes and R. S. Popovic, "Integrated Semiconductor Magnetic Field Sensors," *Proc. IEEE*, **74**, 1107 (1986).
9. P. T. Moseley, "Materials Selection for Semiconductor Gas Sensors," *Sensors Actuators B*, **6**, 149 (1992).
10. D. Kohl, "Function and Applications of Gas Sensors," *J. Phys. D: Appl. Phys.*, **34**, R125 (2001).
11. P. Bergveld, "Development of an Ion-Sensitive Solid-State Device for Neurophysiological Measurements," *IEEE Trans. Biom. Eng.*, **MBE-17**, 70 (1970).
12. P. Bergveld, "Development, Operation, and Application of the Ion-Sensitive Field-Effect Transistor as a Tool for Electrophysiology," *IEEE Trans. Biom. Eng.*, **MBE-19**, 342 (1972).
13. T. Matsuo and K. D. Wise, "An Integrated Field-Effect Electrode for Biopotential Recording," *IEEE Trans. Biom. Eng.*, **MBE-21**, 485 (1974).
14. I. Lundstrom, M. Armgarth, and L. Petersson, "Physics with Catalytic Metal Gate Chemical Sensors," *Crit. Rev. Solid State Mater. Sci.*, **15**, 201 (1989).
15. J. Cooper and T. Cass, *Biosensors: A Practical Approach*, Oxford University, Oxford, 2004.

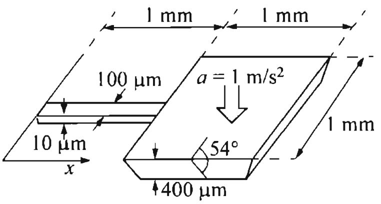
PROBLEMS

1. Derive Eq. 12.

2. A transistor, used as a temperature sensing element, is located at the surface of a silicon chip of 0.5 mm thick. Its junction has an area of 25 µm×25 µm. The transistor is biased with a current of 10 µA, at a collector-emitter voltage of 0.6 V. What is the error in the temperature measurement caused by self-heating of the transistor?
   [*Hint*: Simplify the problem by assuming radial heat flow in a half-sphere. The thermal resistance between concentric spherical surfaces is $R_{th} = (1/4\pi\kappa)[(1/r_1) - (1/r_2)]$ where κ is the thermal conductivity of silicon, 1.5 W/cm-K.]

3. For a silicon strain gauge with a doping of 10^{20} cm^{-3}, find its longitudinal piezoresistive coefficient at 25°C.

4. Take an accelerometer with an inertial mass suspended by one silicon beam. The dimensions are shown in the right figure. Assume a rectangular cross-section of the beam. Calculate the mass of the movable electrode and the stress on the top surface of the suspension beam, as a function of x, when an acceleration of 100 cm/s^2 is applied. Assume there is no gravity, and neglect the mass of the beam. (Density of silicon: 2.33 g/cm^3).



   [*Hint*: The stress at the surface of the beam is given by $6\,M/h^2$ where M is the bending moment = Force×(1 − x), where x is in mm and h is the thickness of the beam.]

5. (a) For an interdigital transducer, if the velocity of SAW propagation is 3.1×10^5 cm/s, and the operating frequency is 840 MHz, find the finger period p.
   (b) The transducer structure is ZnO/SiO$_2$/Si. Let $Kh_{\mathrm{SiO_2}} = 1$ and $Kh_{\mathrm{ZnO}} = 0.3$ to achieve temperature stability where $K \equiv 2\pi/p$. Find the thicknesses $h_{\mathrm{SiO_2}}$ and h_{ZnO} for the SiO$_2$ and ZnO layers.

6. Consider a Hall plate structure similar to Fig. 14 but with opposite dopings (e.g., the epitaxial layer is n-type, and the surrounding and the bottom layers are p-type). Assume t = 10 µm, L = 600 µm, W = 200 µm, a sheet resistance $R_\square$ of 1000 Ω/□, a supply current of I = 10 mA, and a magnetic induction $\mathscr{B}$ = 100 Gauss, find (a) the Hall coefficient, (b) the Hall voltage, and (c) the Hall angle.

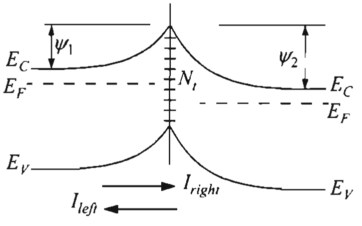
7. (a) Derive the expression for the Hall coefficient R_H for ambipolar current flow.
   (*Hint*: When the magnetic induction vector $\mathscr{B}$ is perpendicular to the electric field $\mathscr{E}$, the electric current density is given by $J_n(\mathscr{B}) = \sigma_{n\mathscr{B}}(\mathscr{E} + \mu_n^*\mathscr{B}\times\mathscr{E})$ where $\sigma_{n\mathscr{B}} = \sigma_n[1 + (\mu_n^*\mathscr{B})^2]^{-1}$, σ_n is the conductivity, μ_n^* is the Hall mobility = $r_n\mu_n$ = r_n×drift mobility.)
   (b) A silicon plate is doped with phosphorus and boron. $N_D = 4.0\times10^{12}$ cm^{-3}, $N_A = 4.1\times10^{12}$ cm^{-3}, $r_n = 1.15$, $r_p = 0.7$, $\mu_p = 0.047/T$, $\mu_n = 0.138/T$. What is the value of R_H?

8. Consider a silicon-based ISFET with L = 1 µm, W = 10 µm, $N_A = 5\times10^{16}$ cm^{-3}, μ_n = 800 cm^2/V-s, and $C_i = 3.45\times10^{-7}$ F/cm^2. For the electrolyte contact to the insulator, we have ϕ_{sol} = 5.30 V, ψ_i = 0.3 V and ψ_{sol} = 0.2 V. Find the current in the saturation region with V_G = 5 V.

9. Under the assumption that there are enough empty surface states so that Richardson's equation $J = AT^2\exp[-(q\psi_s + E_C - E_F)/kT]$ with A = 120 A/cm^2-K^2 is valid for electron transfer to the surface, determine the rate of capture of electrons when a depletion layer is present. As ψ_s becomes more negative, the rate of electron capture as described by the Richardson equation becomes lower and lower. If equilibrium must be reached in less than

10 s for a practical sensor, estimate the allowable limit of band bending. For simplicity, assume as the criterion that the rate of electron capture at equilibrium as given by Richardson's equation must suffice to transfer the surface state charge N_s in 10 s. Assume the temperature is 300 K, $E_C - E_F = 0.15$ eV, the donor density is 10^{17} cm^{-3} and ε_s is 10^{-12} F/cm.

10. Refer to Fig. 22 and the right figure, derive an expression for the resistance of an n-type sample of length L, area W^2 as a function of N_t, where N_t is the density of charge trapped at a grain boundary. Assume there is only one grain boundary extending across the sample, located at L/W^2. Use Richardson's equation from Problem 9. Assume the applied voltage is small.



Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

APPENDIXES

A. List of Symbols
B. International System of Units
C. Unit Prefixes
D. Greek Alphabet
E. Physical Constants
F. Properties of Important Semiconductors
G. Properties of Si and GaAs
H. Properties of SiO_2 and Si_3N_4

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

Appendix A

List of Symbols

Symbol	Description	Unit
a	Lattice constant	Å
A	Area	cm^2
A	Effective Richardson constant for free electron	A/cm^2-K^2
A^*, A^{**}	Effective Richardson constant	A/cm^2-K^2
B	Bandwidth	Hz
$\mathscr{B}$	Magnetic induction	Wb/cm^2, V-s/cm^2
c	Speed of light in vacuum	cm/s
c_s	Speed of sound	cm/s
C_d	Diffusion capacitance per area	F/cm^2
C_D	Depletion-layer capacitance per area	F/cm^2
C_{FB}	Capacitance per area at flat-band	F/cm^2
C_i	Insulator capacitance per area	F/cm^2
C_{it}	Interface-trap capacitance per area	F/cm^2
C_{ox}	Oxide capacitance per area	F/cm^2
C_v	Specific heat	J/g-K
C'	Capacitance	F
d, d_{ox}	Oxide thickness	cm
d_i	Insulator thickness	cm
D	Diffusion coefficient	cm^2/s
D_a	Ambipolar diffusion coefficient	cm^2/s
D_{it}	Interface-trap density	cm^{-2}-eV^{-1}
D_n	Diffusion coefficient for electrons	cm^2/s
D_p	Diffusion coefficient for holes	cm^2/s
$\mathscr{D}$	Electric displacement	C/cm^2
E	Energy	eV
E_a	Activation energy	eV
E_A	Ionization energy for acceptors	eV
E_C	Bottom edge of conduction band	eV
E_D	Ionization energy for donors	eV
E_F	Fermi level	eV

Symbol	Description	Unit
E_{Fm}	Metal Fermi level	eV
E_{Fn}	Quasi-Fermi (imref) level for electrons	eV
E_{Fp}	Quasi-Fermi (imref) level for holes	eV
E_g	Energy gap	eV
E_i	Intrinsic Fermi level	eV
E_p	Optical-phonon energy	eV
E_t	Trap energy level	eV
E_V	Top edge of valence band	eV
$\mathscr{E}$	Electric field	V/cm
$\mathscr{E}_c$	Critical electric field	V/cm
$\mathscr{E}_m$	Maximum electric field	V/cm
f	Frequency	Hz
$f_{\max}$	Maximum frequency of oscillation (unilateral gain is unity)	Hz
f_T	Cutoff frequency	Hz
F	Fermi-Dirac distribution function	—
$F_{1/2}$	Fermi-Dirac integral	—
F_C	Fermi-Dirac distribution function for electrons	—
F_F	Fill factor	—
F_V	Fermi-Dirac distribution function for holes	—
g_m	Transconductance	S
g_{mi}	Transconductance, intrinsic	S
g_{mx}	Transconductance, extrinsic	S
G	Conductance	S
G_a	Gain	—
G_e	Generation rate	$\text{cm}^{-3}\text{-s}^{-1}$
G_n	Electron generation rate	$\text{cm}^{-3}\text{-s}^{-1}$
G_p	Hole generation rate	$\text{cm}^{-3}\text{-s}^{-1}$
G_P	Power gain	—
G_{th}	Thermal generation rate	$\text{cm}^{-3}\text{-s}^{-1}$
h	Planck constant	J-s
h_{fb}	Small-signal common-base current gain, $= \alpha$	—
h_{FB}	Common-base current gain, $= \alpha_0$	—
h_{fe}	Small-signal common-emitter current gain, $= \beta$	—
h_{FE}	Common-emitter current gain, $= \beta_0$	—
$\hbar$	Reduced Planck constant, $h/2\pi$	J-s
$\mathscr{H}$	Magnetic field	A/cm
i	Intrinsic (undoped) material	—

Symbol	Description	Unit
I	Current	A
I_0	Saturation current	A
I_F	Forward current	A
I_h	Holding current	A
I_n	Electron current	A
I_p	Hole current	A
I_{ph}	Photocurrent	A
I_{re}	Recombination current	A
I_R	Reverse current	A
I_{sc}	Short-circuit current in response to light	A
J	Current density	A/cm^2
J_0	Saturation-current density	A/cm^2
J_F	Forward-current density	A/cm^2
J_{ge}	Generation-current density	A/cm^2
J_n	Electron-current density	A/cm^2
J_p	Hole-current density	A/cm^2
J_{ph}	Photocurrent density	A/cm^2
J_{re}	Recombination-current density	A/cm^2
J_R	Reverse-current density	A/cm^2
J_{sc}	Short-circuit-current density	A/cm^2
J_t	Tunneling-current density	A/cm^2
J_T	Threshold-current density	A/cm^2
k	Boltzmann constant	J/K
k	Wave vector	cm^{-1}
k_e	Extinction coefficient, imaginary part of index of refraction	—
k_{ph}	Phonon wave number vector	cm^{-1}
K	Dielectric constant, $\varepsilon/\varepsilon_0$	—
K_i	Dielectric constant of insulator	—
K_{ox}	Dielectric constant of oxide	—
K_s	Dielectric constant of semiconductor	—
L	Length	cm
L	Inductance	H
L_a	Ambipolar diffusion length	cm
L_d	Diffusion length	cm
L_D	Debye length	cm
L_n	Diffusion length of electrons	cm
L_p	Diffusion length of holes	cm

Symbol	Description	Unit
m_0	Electron rest mass	kg
m^*	Effective mass	kg
m_c^*	Conductivity effective mass	kg
m_{ce}^*	Conductivity effective mass for electrons	kg
m_{ch}^*	Conductivity effective mass for holes	kg
m_{de}^*	Density-of-state effective mass for electrons	kg
m_{dh}^*	Density-of-state effective mass for holes	kg
m_e^*	Electron effective mass	kg
m_h^*	Hole effective mass	kg
m_{hh}^*	Effective mass for heavy hole	kg
m_l^*	Longitudinal effective mass for electron	kg
m_{lh}^*	Effective mass for light hole	kg
m_t^*	Transverse effective mass for electron	kg
M	Multiplication factor	—
M_C	Number of equivalent minima in the conduction band	—
M_n	Multiplication factor of electrons	—
M_p	Multiplication factor of holes	—
n	Concentration of free electron	cm^{-3}
n	Of *n*-type semiconductor (with donor impurity)	—
n_i	Intrinsic carrier concentration	cm^{-3}
n_n	Electron concentration in *n*-type semiconductor (majority carriers)	cm^{-3}
n_{no}	n_n in thermal equilibrium	cm^{-3}
n_p	Electron concentration in *p*-type semiconductor (minority carriers)	cm^{-3}
n_{po}	n_p in thermal equilibrium	cm^{-3}
n_r	Real part of refractive index	—
$\bar{n}$	Complex refractive index, $= n_r + ik_e$	—
N	Doping concentration	cm^{-3}
N	Density of states	eV^{-1}-cm^{-3}
N_A	Acceptor impurity concentration	cm^{-3}
N_A^-	Ionized acceptor impurity concentration	cm^{-3}
N_b'	Gummel number	cm^{-2}
N_C	Effective density of states in conduction band	cm^{-3}
N_D	Donor impurity concentration	cm^{-3}
N_D^+	Ionized donor impurity concentration	cm^{-3}
N_t	Bulk-trap concentration	cm^{-3}
N_V	Effective density of states in valence band	cm^{-3}
N^*	Density per area	cm^{-2}
N_{it}^*	Interface-trap density per area	cm^{-2}
N_{st}^*	Surface-trap density per area	cm^{-2}

Symbol	**Description**	**Unit**
p	Concentration of free hole	cm^{-3}
p	Of p-type semiconductor (with acceptor impurity)	—
p	Momentum	J-s/cm
p_n	Hole concentration in n-type semiconductor (minority carriers)	cm^{-3}
p_{no}	p_n in thermal equilibrium	cm^{-3}
p_p	Hole concentration in p-type semiconductor (majority carriers)	cm^{-3}
p_{po}	p_p in thermal equilibrium	cm^{-3}
P	Pressure	N/cm^2
P	Power	W
P_{op}	Optical power density or intensity	W/cm^2
P_{opt}	Total optical power	W
q	Unit electronic charge, = 1.6×10^{-19} C, absolute value	C
Q	Quality factor of capacitor and inductor	—
Q	Charge density	C/cm^2
Q_D	Space-charge density in depletion region	C/cm^2
Q_f	Fixed-oxide-charge density	C/cm^2
Q_{it}	Interface-trap-charge density	C/cm^2
Q_m	Mobile-ionic-charge density	C/cm^2
Q_{ot}	Oxide-trapped-charge density	C/cm^2
r_F	Dynamic forward resistance	Ω
r_H	Hall factor	—
r_R	Dynamic reverse resistance	Ω
R	Reflection of light	—
R	Resistance	Ω
R_c	Specific contact resistance	Ω-cm^2
R_{co}	Contact resistance	Ω
R_{CG}	Coupling ratio of floating gate	—
R_e	Recombination rate	cm^{-3}-s^{-1}
R_{ec}	Recombination coefficient	cm^3/s
R_H	Hall coefficient	cm^3/C
R_L	Load resistance	Ω
R_{nr}	Nonradiative recombination rate	cm^{-3}-s^{-1}
R_r	Radiative recombination rate	cm^{-3}-s^{-1}
$R_\square$	Sheet resistance per square	$\Omega/\square$
$\mathscr{R}$	Responsivity	A/W
S	Strain	—
S	Subthreshold swing	V/decade of current

Symbol	Description	Unit
S_n	Surface recombination velocity for electrons	cm/s
S_p	Surface recombination velocity for holes	cm/s
t	Time	s
t_r	Transit time	s
T	Absolute temperature	K
T	Stress	N/cm^2
T	Transmission of light	—
T_e	Electron temperature	K
T_t	Tunneling probability	—
U	Net recombination/generation rate, $U = R - G$.	cm^{-3}-s^{-1}
v	Carrier velocity	cm/s
v_d	Drift velocity	cm/s
v_g	Group velocity	cm/s
v_n	Electron velocity	cm/s
v_p	Hole velocity	cm/s
v_{ph}	Phonon velocity	cm/s
v_s	Saturation velocity	cm/s
v_{th}	Thermal velocity	cm/s
V	Applied voltage	V
V_A	Early voltage	V
V_B	Breakdown voltage	V
V_{BCBO}	Collector-base open-emitter breakdown voltage	V
V_{BCEO}	Collector-emitter open-base breakdown voltage	V
V_{BS}	Back-substrate voltage	V
V_{CC}, V_{DD}	Supply voltage	V
V_F	Forward bias	V
V_{FB}	Flat-band voltage	V
V_h	Holding voltage	V
V_H	Hall voltage	V
V_{oc}	Open-circuit voltage in response to light	V
V_P	Pinch-off voltage	V
V_{PT}	Punch-through voltage	V
V_R	Reverse bias	V
V_T	Threshold voltage	V
W	Thickness	cm
W_B	Base thickness	cm
W_D	Depletion width	cm

Symbol	Description	Unit
W_{Dm}	Maximum depletion width	cm
W_{Dn}	Depletion width in n-type material	cm
W_{Dp}	Depletion width in p-type material	cm
x	Distance or thickness	cm
Y	Young's modulus, modulus of elasticity	N/cm^2
Z	Impedance	Ω
α	Optical absorption coefficient	cm^{-1}
α	Small-signal common-base current gain, $= h_{fb}$	—
α	Ionization coefficient	cm^{-1}
α_0	Common-base current gain, $= h_{FB}$	—
α_n	Ionization coefficient for electrons	cm^{-1}
α_p	Ionization coefficient for holes	cm^{-1}
α_T	Base transport factor	—
β	Small-signal common-emitter current gain, $= h_{fe}$	—
β_0	Common-emitter current gain, $= h_{FE}$	—
β_{th}	Reciprocal of thermal potential, $= q/kT$	V^{-1}
γ	Emitter injection efficiency	—
Δn	Excess electron concentration beyond equilibrium	cm^{-3}
Δp	Excess hole concentration beyond equilibrium	cm^{-3}
ε	Permittivity	F/cm, C/V-cm
ε_0	Permittivity of vacuum	F/cm, C/V-cm
ε_i	Permittivity of insulator	F/cm, C/V-cm
ε_{ox}	Permittivity of oxide	F/cm, C/V-cm
ε_s	Permittivity of semiconductor	F/cm, C/V-cm
η	Quantum efficiency	—
η	Ideality factor of rectifier under forward bias	—
η_{ex}	External quantum efficiency	—
η_{in}	Internal quantum efficiency	—
θ	Angle	rad, °

Symbol	Description	Unit
κ	Thermal conductivity	W/cm-K
λ	Wavelength	cm
λ_m	Mean free path	cm
λ_{ph}	Phonon mean free path	cm
μ	Drift mobility ($\equiv v/\mathscr{E}$)	cm^2/V-s
μ	Permeability	H/cm
μ_0	Permeability in vacuum	H/cm
μ_d	Differential mobility ($\equiv dv/d\mathscr{E}$)	cm^2/V-s
μ_H	Hall mobility	cm^2/V-s
μ_n	Electron mobility	cm^2/V-s
μ_p	Hole mobility	cm^2/V-s
ν	Frequency of light	Hz, s^{-1}
ν	Poisson's ratio	—
ν	Lightly doped *n*-type material	—
π	Lightly doped *p*-type material	—
ρ	Resistivity	Ω-cm
ρ	Charge density	C/cm^3
σ	Conductivity	$S\text{-}cm^{-1}$
σ	Capture cross-section	cm^2
σ_n	Capture cross-section for electrons	cm^2
σ_p	Capture cross-section for holes	cm^2
τ	Carrier lifetime	s
τ_a	Ambipolar carrier lifetime	s
τ_A	Auger lifetime	s
τ_e	Energy relaxation time	s
τ_g	Carrier generation lifetime	s
τ_m	Mean free time in scattering	s
τ_n	Carrier lifetime for electrons	s
τ_{nr}	Carrier lifetime due to nonradiative recombination	s
τ_p	Carrier lifetime for holes	s
τ_r	Carrier lifetime due to radiative recombination	s
τ_R	Dielectric relaxation time	s
τ_s	Storage time	s
τ_t	Transit time	s

Symbol	Description	Unit
ϕ	Work function or barrier height	V
ϕ_B	Barrier height	V
ϕ_{Bn}	Schottky barrier height on *n*-type semiconductor	V
ϕ_{Bp}	Schottky barrier height on *p*-type semiconductor	V
ϕ_m	Metal work function	V
ϕ_{ms}	Work-function difference between metal and semiconductor, $\phi_m - \phi_s$	V
ϕ_n	Fermi potential from conduction-band edge in *n*-type semiconductor, $(E_C - E_F)/q$. Negative for degenerate material (see figure)	V
ϕ_p	Fermi potential from valence-band edge in *p*-type semiconductor, $(E_F - E_V)/q$. Negative for degenerate material (see figure)	V
ϕ_s	Semiconductor work function	V
ϕ_{th}	Thermal potential, kT/q	V
Φ	Photon flux	s^{-1}
χ	Electron affinity	V
χ_s	Electron affinity for semiconductor	V
ψ	Wavefunction	—
ψ_{bi}	Built-in potential at equilibrium (always positive)	V
ψ_B	Fermi level from intrinsic Fermi level, $\|E_F - E_i\|/q$, in bulk	V
ψ_{Bn}	ψ_B in *n*-type material (see figure)	V
ψ_{Bp}	ψ_B in *p*-type material (see figure)	V
ψ_i	Semiconductor potential, $-E_i/q$	V

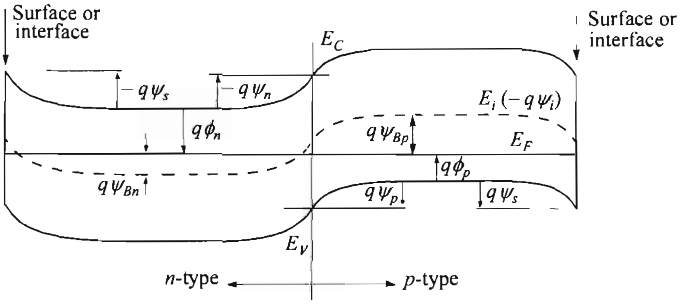


Fig. 1 Symbols and definitions for semiconductor potentials. Note that surface potentials are with respect to the bulk, and are positive when the band bends downwards. ϕ_n and ϕ_p are negative when E_F are outside the bandgap (degenerate).

Symbol	**Description**	**Unit**
ψ_n	Potential at *n*-type boundary with respect to *n*-type bulk (band bending in *n*-type material, positive when bending down in the energy-band diagram) (see figure)	V
ψ_p	Potential at *p*-type boundary with respect to *p*-type bulk (band bending in *p*-type material, positive when bending down in the energy-band diagram) (see figure)	V
ψ_s	Surface potential with respect to bulk (band bending, positive when bending down in the energy-band diagram) (see figure)	V
ω	Angular frequency, $= 2\pi f$ or $2\pi\nu$	Hz

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

Appendix B

International System of Units

Quantity	**Unit**	**Symbol**	**Dimensions**
Length	meter*	m*	
Mass	kilogram	kg	
Time	second	s	
Temperature	kelvin	K	
Current	ampere	A	C/s
Frequency	hertz	Hz	s^{-1}
Force	newton	N	kg-m/s^2, J/m
Pressure, stress	pascal	Pa	N/m^2
Energy	joule*	J*	N-m, W-s
Power	watt	W	J/s, V-A
Electric charge	coulomb	C	A-s
Potential	volt	V	J/C, W/A
Conductance	siemens	S	A/V, 1/Ω
Resistance	ohm	Ω	V/A
Capacitance	farad	F	C/V
Magnetic flux	weber	Wb	V-s
Magnetic induction	tesla	T	Wb/m^2
Inductance	henry	H	Wb/A

* It is more common in the semiconductor field to use cm for length and eV for energy. (1 cm = 10^{-2} m, 1 eV = 1.6×10^{-19} J)

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

Appendix C

Unit Prefixes [#]

Multiple	Prefix	Symbol
10^{18}	exa	E
10^{15}	peta	P
10^{12}	tera	T
10^{9}	giga	G
10^{6}	mega	M
10^{3}	kilo	k
10^{2}	hecto	h
10	deka	da
10^{-1}	deci	d
10^{-2}	centi	c
10^{-3}	milli	m
10^{-6}	micro	μ
10^{-9}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a

[#] Adopted by International Committee on Weights and Measures. (Compound prefixes should not be used; e.g., use p instead of μμ.)

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

Appendix D

Greek Alphabet

Letter	**Lower Case**	**Upper Case**
Alpha	α	A
Beta	β	B
Gamma	γ	Γ
Delta	δ	Δ
Epsilon	ε	E
Zeta	ζ	Z
Eta	η	H
Theta	θ	Θ
Iota	ι	I
Kappa	κ	K
Lambda	λ	Λ
Mu	μ	M
Nu	ν	N
Xi	ξ	Ξ
Omicron	ο	O
Pi	π	Π
Rho	ρ	P
Sigma	σ	Σ
Tau	τ	T
Upsilon	υ	Y
Phi	ϕ	Φ
Chi	χ	X
Psi	ψ	Ψ
Omega	ω	Ω

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

Appendix E

Physical Constants

Quantity	Symbol	Value
Atmospheric pressure		1.01325×10^5 N/cm^2
Avogadro constant	N_{AV}	6.02204×10^{23} mol^{-1}
Bohr radius	a_B	0.52917 Å
Boltzmann constant	k	1.38066×10^{-23} J/K (R/N_{AV})
		8.6174×10^{-5} eV/K
Electron rest mass	m_0	9.1095×10^{-31} kg
Electron-volt energy	eV	1 eV = 1.60218×10^{-19} J
Elementary charge	q	1.60218×10^{-19} C
Gas constant	R	1.98719 cal/mol-K
Magnetic flux quantum ($h/2q$)		2.0678×10^{-15} Wb
Permeability in vacuum	μ_0	1.25663×10^{-8} H/cm ($4\pi\times10^{-9}$)
Permittivity in vacuum	ε_0	8.85418×10^{-14} F/cm ($1/\mu_0 c^2$)
Planck constant	h	6.62617×10^{-34} J-s
		4.1357×10^{-15} eV-s
Proton rest mass	M_p	1.67264×10^{-27} kg
Reduced Planck constant ($h/2\pi$)	$\hbar$	1.05458×10^{-34} J-s
		6.5821×10^{-16} eV-s
Speed of light in vacuum	c	2.99792×10^{10} cm/s
Thermal voltage at 300 K	kT/q	0.0259 V

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

Appendix F
Properties of Important Semiconductors

Semiconductor			Crystal Struct.	Lattice Const. at 300 K (Å)	Bandgap (eV)		Band	Mobility at 300 K (cm^2/V-s)		Effective Mass		$\varepsilon_s/\varepsilon_0$
					300 K	0 K		μ_n	μ_p	m_n^*/m_0	m_p^*/m_0	
	C	Carbon (diamond)	D	3.56683	5.47	5.48	I	1,800	1,200	0.2	0.25	5.7
	Ge	Germanium	D	5.64613	0.66	0.74	I	3,900	1,900	$1.64^{l}, 0.082^{t}$	$0.04^{lh}, 0.28^{hh}$	16.0
	Si	Silicon	D	5.43102	1.12	1.17	I	1,450	500	$0.98^{l}, 0.19^{t}$	$0.16^{lh}, 0.49^{hh}$	11.9
IV-IV	SiC	Silicon carbide	W	a=3.086,c=15.117	2.996	3.03	I	400	50	0.60	1.00	9.66
III-V	AlAs	Aluminum arsenide	Z	5.6605	2.36	2.23	I	180		0.11	0.22	10.1
	AlP	Aluminum phosphide	Z	5.4635	2.42	2.51	I	60	450	0.212	0.145	9.8
	AlSb	Aluminum antimonide	Z	6.1355	1.58	1.68	I	200	420	0.12	0.98	14.4
	BN	Boron nitride	Z	3.6157	6.4		I	200	500	0.26	0.36	7.1
	”	”	W	a=2.55,c=4.17	5.8		D			0.24	0.88	6.85
	BP	Boron phosphide	Z	4.5383	2.0		I	40	500	0.67	0.042	11
	GaAs	Gallium arsenide	Z	5.6533	1.42	1.52	D	8,000	400	0.063	$0.076^{lh}, 0.5^{hh}$	12.9
	GaN	Gallium nitride	W	a=3.189,c=5.182	3.44	3.50	D	400	10	0.27	0.8	10.4
	GaP	Gallium phosphide	Z	5.4512	2.26	2.34	I	110	75	0.82	0.60	11.1
	GaSb	Gallium antimonide	Z	6.0959	0.72	0.81	D	5,000	850	0.042	0.40	15.7
	InAs	Indium arsenide	Z	6.0584	0.36	0.42	D	33,000	460	0.023	0.40	15.1
	InP	Indium phosphide	Z	5.8686	1.35	1.42	D	4,600	150	0.077	0.64	12.6
	InSb	Indium antimonide	Z	6.4794	0.17	0.23	D	80,000	1,250	0.0145	0.40	16.8
II-VI	CdS	Cadmium sulfide	Z	5.825	2.5		D			0.14	0.51	5.4
	”	”	W	a=4.136,c=6.714	2.49		D	350	40	0.20	0.7	9.1
	CdSe	Cadmium selenide	Z	6.050	1.70	1.85	D	800		0.13	0.45	10.0
	CdTe	Cadmium telluride	Z	6.482	1.56		D	1,050	100			10.2
	ZnO	Zinc oxide	R	4.580	3.35	3.42	D	200	180	0.27		9.0
	ZnS	Zinc sulfide	Z	5.410	3.66	3.84	D	600		0.39	0.23	8.4
	”	”	W	a=3.822,c=6.26	3.78		D	280	800	0.287	0.49	9.6
IV-VI	PbS	Lead sulfide	R	5.9362	0.41	0.286	I	600	700	0.25	0.25	17.0
	PbTe	Lead telluride	R	6.4620	0.31	0.19	I	6,000	4,000	0.17	0.20	30.0

D = Diamond, W = Wurtzite, Z = Zincblende, R = Rock salt. I, D = Indirect, direct bandgap. *l,t,lh,hh* = Longitudinal,transverse,light-hole,heavy-hole effective mass.

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

Appendix G

Properties of Si and GaAs

Property	**Si**	**GaAs**
Atom density (cm^{-3})	5.02×10^{22}	4.43×10^{22}
Atomic weight	28.09	144.64
Crystal structure	Diamond	Zincblende
Density (g/cm^3)	2.329	5.317
Lattice constant (Å)	5.43102	5.6533
Dielectric constant	11.9	12.9
Electron affinity χ (V)	4.05	4.07
Energy gap (eV)	1.12 (indirect)	1.42 (direct)
Effective density of states in conduction band, N_C (cm^{-3})	2.8×10^{19}	4.7×10^{17}
Effective density of states in valence band, N_V (cm^{-3})	2.65×10^{19}	7.0×10^{18}
Intrinsic carrier concentration n_i (cm^{-3})	9.65×10^{9}	2.1×10^{6}
Effective mass (m^*/m_0) Electrons	$m_l^* = 0.98$ $m_t^* = 0.19$	0.063
Holes	$m_{lh}^* = 0.16$ $m_{hh}^* = 0.49$	$m_{lh}^* = 0.076$ $m_{hh}^* = 0.50$
Drift mobilities (cm^2/V-s) Electrons μ_n	1,450	8,000
Holes μ_p	500	400
Saturation velocity (cm/s)	1×10^{7}	7×10^{6}
Breakdown field (V/cm)	$2.5–8\times10^{5}$	$3–9\times10^{5}$
Minority-carrier lifetime (s)	$\approx10^{-3}$	$\approx10^{-8}$
Index of refraction	3.42	3.3
Optical-phonon energy (eV)	0.063	0.035
Melting point (°C)	1414	1240
Linear coefficient of thermal expansion $\Delta L/L\Delta T$ (°C^{-1})	2.59×10^{-6}	5.75×10^{-6}
Thermal conductivity (W/cm-K)	1.56	0.46
Thermal diffusivity (cm^2/s)	0.9	0.31
Specific heat (J/g-°C)	0.713	0.327
Heat capacity (J/mol-°C)	20.07	47.02
Young's modulus (GPa)	130	85.5

Note: All properties at room temperature.

Physics of Semiconductor Devices, 3rd Edition
by S. M. Sze and Kwok K. Ng
Copyright © 2007 John Wiley & Sons, Inc.

Appendix H

Properties of SiO_2 and Si_3N_4

Property	**SiO_2**	**Si_3N_4**
Structure	Amorphous	Amorphous
Density (g/cm^3)	2.27	3.1
Dielectric constant	3.9	7.5
Dielectric strength (V/cm)	$\approx 10^7$	$\approx 10^7$
Electron affinity, χ (eV)	0.9	
Energy gap, E_g (eV)	9	≈ 5
Infrared absorption band (μm)	9.3	11.5–12.0
Melting point (°C)	≈ 1700	
Molecular density (cm^{-3})	2.3×10^{22}	
Molecular weight	60.08	
Refractive index	1.46	2.05
Resistivity (Ω-cm)	10^{14}–10^{16}	$\approx 10^{14}$
Specific heat (J/g-°C)	1.0	
Thermal conductivity (W/cm-K)	0.014	
Thermal diffusivity (cm^2/s)	0.006	
Thermal expansion coef., linear (°C^{-1})	5.0×10^{-7}	

Note: All properties at room temperature.

Copyright © 2007 John Wiley & Sons, Inc.

Index

1/*f* noise 118, 271, 435, 666, 763

Abrupt junction 80
Absorption 622, 623, 625, 626
Absorption coefficient 51, 52, 65, 622, 664, 671, 680, 682, 692, 718, 726, 730, 732, 733
ac trigger diode 577
Acceleration 752, 753
Accelerometer 753
Acceptor impurity 21, 22, 80
Acceptor interface trap 140, 141, 214
Acceptor level 21
Accumulation 200, 201
Accumulation-layer mode 528
Acoustic mode 50
Acoustic phonon 28, 36
Acoustic sensor 754
Activation energy 173, 174
Active pixel sensor 712
ADC 712
Air mass 720
AM noise 534
AM0 720
AM1 720
Ambipolar diffusion coefficient 93, 566, 593
Ambipolar diffusion length 567
Ambipolar lifetime 566
Amplification 393
Amplifier 407, 548
Amplitude deviations (AM) noise 534
Analog gain 275
Analog transistor 586, 590
Analog-to-digital converter 712
Anisotype heterojunction 124
Antireflection coating 682, 690, 726, 731
Antiresonant 480
APD 683
  heterojunction 693
  heterostructure 692
  metal-semiconductor 689
  Schottky-barrier 691, 692
Apodisation 754
APS 712
Arbitrary doping profile 88
Auger coefficient 565
Auger effect 258
Auger lifetime 250
Auger process 40, 565, 566, 603
Auger recombination 249, 250, 568
Automatic gain control 712
Avalanche 210, 444, 447, 476, 485, 486, 607, 685
Avalanche breakdown 106, 112, 172, 184, 255, 278, 346, 356, 390, 467, 472, 474, 484, 493, 499, 514, 525, 551, 577, 584, 608
Avalanche delay 466, 497
Avalanche emission 603
Avalanche excitation 608
Avalanche multiplication 79, 96, 102, 104, 120, 231, 232, 250, 390, 447, 479, 482, 492, 552, 557, 671, 690, 691
Avalanche multiplication factor 436
Avalanche phase delay 487, 504
Avalanche photodiode 633, 671, 681, 683, 694
  heterojunction 692
Avalanche region 466, 467, 469, 471, 472, 474, 477, 478, 479, 480, 481, 485, 489, 491, 493
Avalanche width 470, 471, 474, 492

Back scattering 312
Back surface field 731
Background charge 708
Backward diode 435
Ball alloy 431
Ballasting resistor 277
Ballistic transport 37, 49, 294, 303, 309, 310, 312, 386, 453, 590

Band structure 12
Bandgap (also see energy gap) 2, 12, 13, 15, 20, 22, 25, 37, 40, 44, 51, 52, 53, 96, 102, 104, 107, 110, 124, 125, 126, 127, 136, 143, 144, 168, 172, 178, 189, 213, 215, 221, 223, 225, 258, 259, 282, 283, 284, 285, 286, 287, 295, 374, 398, 401, 408, 423, 430, 431, 450, 454, 497, 514, 549, 603, 605, 606, 607, 611, 612, 613, 618, 622, 625, 626, 629, 630, 633, 636, 640, 650, 664, 667, 680, 692, 693, 714, 721, 722, 724, 730, 734, 736
  direct 7, 15, 40, 42, 52, 122, 421, 604, 605, 608, 612, 614, 621, 630, 633, 680, 692, 716
  indirect 15, 42, 427, 604, 605, 608, 611, 614, 630, 633, 716, 732
Bandgap narrowing 258, 283, 568
Band-to-band excitation 176, 682, 713
Band-to-band photoexcitation 672
Band-to-band transition 40, 667
Band-to-band tunneling 104, 120, 419
BARITT diode 2, 495, 497, 505
Barrier height 47, 128, 135, 136, 141, 142, 148, 153, 154, 156, 159, 174, 175, 178, 180, 188, 228, 342, 357, 377, 439, 440, 442, 449, 450, 454, 499, 531, 532, 588, 589, 594, 664, 734
Barrier lowering 148
Barrier-height adjustment 150
Barrier-injection transit-time diode 497
Base charging time 264
Base resistance 251, 261, 262, 275, 276, 282
Base transport factor 248, 252, 338, 452, 553, 585, 696
Base width 253, 259, 261, 275, 452, 561, 562
BAW device 755
BBD 704
bcc reciprocal lattice 11, 12
BCCD 699, 706
Beta particle 553
Bevel angle 553, 554, 555
Bias charge 708
BiCMOS 282
Bidirectional *p-n-p-n* diode 577, 578
Bidirectional thyristor 577
Binding energy 613
Biosensor 768
Bipolar transistor 2, 66, 124, 184, 243, 295, 296, 314, 349, 388, 418, 450, 541, 548, 552, 573, 577, 582, 583, 584, 585, 586, 590, 591, 694, 695, 696, 748, 762
  double-heterojunction 284, 285
  graded-base 284, 285, 286
  heterojunction 127, 267, 282
Bipolar-mode SIT 590
Bistable 549, 550
Black-body radiation 623, 666
Bloch function 12
Bloch theorem 12
Blockade voltage 364
Blocking-voltage gain 594
Body-centered cubic reciprocal lattice 11
Boltzmann approximation 90
Boltzmann distribution 34, 606
Boltzmann relation 91
Boltzmann statistics 19, 81, 127, 233, 623
Bound-to-bound 716, 717
Bound-to-continuum 716, 717
Bound-to-extended 717
Bound-to-miniband 716, 717
Bragg reflection 636, 637
Breakdown 79, 96, 102, 106, 107, 172, 182, 184, 234, 255, 276, 278, 296, 330, 338, 341, 346, 356, 375, 388, 389, 390, 397, 436, 467, 471, 472, 474, 478, 484, 493, 496, 514, 525, 552, 555, 577, 584, 608, 669, 671, 689, 690
Breakdown voltage 104, 106, 107, 110, 111, 112, 114, 120, 123, 154, 172, 231, 255, 257, 258, 267, 276, 278, 280, 285, 338, 388, 389, 409, 436, 447, 467, 469, 470, 472, 473, 484, 489, 499, 549, 550, 551, 552, 553, 555, 558, 559, 569, 572, 577, 583, 590, 683, 689
Breakover 550, 551, 557, 559, 562, 571, 578
Breakover diode 583
Brillouin zone 11, 12, 13, 36, 50, 513
Broken-gap heterojunction 58
BSIT 590
Bucket-brigade device 704
Built-in field 248, 249, 264, 266
Built-in potential 80, 81, 83, 88, 90, 112, 121, 126, 148, 154, 175, 266, 283, 284, 375, 377, 392, 393, 408, 410, 424, 429, 445,

470, 497, 563, 587, 589, 590, 593, 625, 677, 723
Bulk-acoustic-wave device 755
Buried channel 296, 320, 324, 326, 327, 375
Buried-channel CCD 699, 709, 710
Buried-channel FET 392, 587
Buried-channel MOSFET 326, 327, 375
Burnout 489

Cantilever 758
Capacitance 431
  gate 380, 396
  gate-channel 395
  gate-drain 396
  gate-source 385, 392
  input 396
  output 395
  parasitic input 395, 396
Capacitive sensor 757
Capacitor 374, 483
  metal-insulator-semiconductor 1, 197
  metal-oxide-silicon 213
  MIS 1, 197, 437
  MOS 1, 213
Capture cross section 43, 615
Carrier confinement 635, 640
Carrier degeneracy 311
Carrier domain 763
Carrier lifetime 40, 42, 101, 122, 647, 648, 664, 669
Carrier-carrier scattering 566, 568
Carrier-domain magnetic-field sensor 763
Catalytic-metal sensor 768
Cathode short 559, 563, 574, 582, 583, 585
Cathodoluminescence 607
CCD 697, 698, 712
  buried-channel 699, 709, 710
  surface-channel 699, 710
CCD image sensor 698
Channel doping profile 339
Channel length 298, 304, 312, 328, 332, 333, 338, 348, 376, 400
  effective 333, 380
Channel potential 304
Channel resistance 395, 397
Channel width 298, 376, 399
Channel-length modulation 333
Charge neutrality 22, 93, 201, 593
Charge packet 699, 700, 703, 710
Charge retention 352
Charge sharing 331
Charge storage 682
Charge to breakdown 235
Charge transfer 704, 705, 709, 711
Charge-control model 116
Charge-couple device 697
Charge-coupled image sensor 697
Charge-injection device 701
Charge-injection transistor 541
Charge-sheet model 302
Charge-storage diode 122
Charge-transfer device 698
Charge-transfer image sensor 697
Charge-trapping device 351, 357
Chemical sensor 3, 743, 765, 768
Chemically sensitive field-effect transistor 766
Child-Langmuir law 49
CHINT 541
CID 701
Clock frequency 707, 709, 710
Close-packed lattice 8
CMOS 349
CMOS image sensor 711
Coherent 623
Coherent tunneling 456
Collector efficiency 452
Collector-up 285
Color converter 619
Comb transducer 754
COMFET 582, 583
Common-base 244, 254
Common-base current gain 251, 269, 338, 450, 552, 557, 561, 577
Common-collector 244
Common-emitter 244, 252, 254, 256, 258, 271, 278, 279, 285
Common-emitter current gain 252, 253, 263, 269, 452, 697
Commutation 586
Complementary bipolar transistor 281
Complementary MOS (CMOS) 349
Complete elliptic integral of the first kind 716
Concentrator 719, 735
Conductance 30, 431
Conduction band 13

Conductivity effective mass 21, 28
Conductivity mobility 30
Conductivity modulation 276, 277, 583, 593
Conductivity-modulated FET 583
Conductivity-modulated field-effect transistor 582
Confinement factor 629, 630, 640
Constant mobility 303, 306, 307, 308, 312, 326, 379, 381, 383, 384, 385, 391, 392, 405
Constant-current phase 115, 116, 117
Constant-field scaling 329, 330
Constant-voltage scaling 330
Contact potential 136
Continuity equation 63, 93, 101, 116, 246, 254, 259, 518, 561, 646, 648, 726
Conversion efficiency 719, 721, 724, 734
Corbino disc 761
Correction factor 31
Coulomb blockade 361, 362, 363, 365
Coulomb scattering 28
Coulomb-blockade diamond 365
Coulomb-blockade oscillation 361, 363
Coulomb-blockade voltage 365
Coupling ratio 354
Critical angle 616, 628
Critical field 107, 307, 382, 406, 484, 536
Critical thickness 57, 408
Crystal 8
Crystal plane 8
Current crowding 261, 276, 282
Current equation 62, 81, 158, 254, 518, 519
Current gain 250, 251, 252, 253, 257, 259, 276, 283, 286, 548, 552, 559, 560, 563, 584, 585, 683
  common-base 251, 269, 338, 450, 552, 557, 561, 577
  common-emitter 252, 253, 263, 269, 452, 697
Current peak 456
Current valley 456
Curvature coefficient 435, 436
Cutoff frequency (f_T) 185, 263, 264, 266, 348, 395, 407, 541, 620
  reactive 433
  resistive 433
Cylindrical *p-n* junction 112
Dark current 664, 666, 670, 682, 683, 684, 687, 692, 696, 699, 700, 701, 709, 712, 714, 718, 719, 722
Dark-line defect 651
Darlington phototransistor 697
Dawson's integral 161
DBR laser 636
de Broglie wavelength 59, 61, 651, 654
Dead space 693
Dead zone 531
Debye length 85, 86, 202, 205, 301, 519, 589
Decay half-life 553
Deceleration emission 603
Deep depletion 209, 229, 231, 445, 448, 699
Deep-level impurity 139, 176
Defect density 430
Deflection mode 763
Degenerate 19, 163, 166, 188, 418, 421, 425, 435, 437, 442, 452
Delay time 646, 648
Density modulation 539, 542
Density of states 17, 61, 62, 423, 424, 425, 514, 653
  joint 606
Density-of-state effective mass 17, 19, 421
Depleted-base transistor 590
Depletion 200, 202, 314, 593
Depletion approximation 49, 81, 83, 88, 148, 377
Depletion charge 302
Depletion layer 84, 85, 106, 136, 216, 295, 374, 376, 377, 392, 472, 551, 672
Depletion mode 296, 326, 327, 349, 376, 404, 587
Depletion region 81, 82, 83, 88, 92, 96, 97, 98, 99, 105, 106, 136, 139, 154, 158, 162, 187, 201, 206, 207, 208, 209, 246, 250, 251, 253, 256, 264, 265, 284, 300, 302, 320, 325, 331, 334, 387, 431, 444, 445, 447, 473, 557, 563, 587, 593, 679
Depletion width 80, 83, 88, 112, 323, 375, 376, 377, 378, 379, 380, 385, 389, 390, 433, 467, 470, 471, 487, 493, 497, 504, 551, 576, 594, 666, 674, 683, 713
Depletion-layer capacitance 85, 90, 315, 475, 481
Depletion-layer charge 90
Depletion-mode 296

Detectivity 667, 668, 670, 671
Detector 743
Deuteron irradiation 44
Device building block 1
DFB laser 636, 637
DH laser 634, 643, 644, 651
DHBT 284, 285
Diac 577, 578
Diamond lattice 8, 13
Diaphragm 752, 758
DIBL 333, 335
Dielectric breakdown 234
Dielectric constant 149, 178, 331, 340
Dielectric relaxation frequency 357, 527
Dielectric relaxation time 148, 519, 524, 527
Differential mobility 514, 519
Differential negative resistance 37
Differential pressure 752
Differential resistance 185, 538
Diffusion 45, 154, 445, 679, 705, 709
Diffusion capacitance 100, 101, 102, 264
Diffusion coefficient 45, 571, 675
  ambipolar 93, 566, 593
Diffusion conductance 100, 101
Diffusion constant 518, 705
Diffusion current 46, 93, 94, 97, 98, 119, 127, 245, 250, 314, 315, 419, 428, 429, 430, 589, 594, 674, 676, 725, 746, 748
Diffusion equation 115, 675
Diffusion length 45, 46, 66, 92, 247, 250, 253, 338, 553, 561, 562, 576, 679, 709, 727, 730
  ambipolar 567
Diffusion potential 81
Diffusion theory 154, 158, 159, 161
Diffusion velocity 589
Diffusivity 45
Digital pixel sensor 712
Digital signal processing 712
Diode
  ac trigger 577
  backward 435
  BARITT 2, 495, 497, 505
  barrier-injection transit-time 497
  bidirectional *p-n-p-n* 577, 578
  breakover 583
  charge-storage 122
  double-barrier 454
Diode (*Cont.*)
  double-drift 467, 485, 486, 494
  Esaki 418
  fast-recovery 122
  gated 112, 338
  Gunn 418, 511
  hi-lo 467
  impact-ionization avalanche transit-time 466
  IMPATT 2, 184, 418, 466, 530, 532, 534, 690
  interband tunnel 418
  light-emitting 3, 601, 608
  lo-hi-lo 467, 532
  metal-insulator-metal tunnel 448
  metal-insulator-semiconductor tunnel 169
  metal-semiconductor 122
  MIM 448
  MIM tunnel 448
  MIS switch 444
  MIS tunnel 169, 437, 439, 442, 444, 768
  Misawa 467, 484
  modified Read 467, 469
  p-i-n 123, 467, 469, 473, 474, 557, 584, 591, 593, 762
  reach-through 497
  Read 467, 469, 471, 472, 477, 482, 484
  real-space-transfer 514, 536
  resonant-tunneling 2, 418, 454
  RST 536, 538
  Schottky 122, 154, 272, 682
  Schottky-barrier 46, 437, 682, 768
  Shockley 444, 550, 569, 571, 578
  single-drift 486, 494
  snapback 122
  step-recovery 122
  tunnel 2, 418, 424, 425, 466, 514
  tunnel-injection transit-time 505
  TUNNETT 2, 504
  two-sided 467
  vacuum 541
  Zener 120
Diode ac switch 577
Diode thermal sensor 746
Dipole 517, 521, 525, 530
Dipole layer 173
Dipole-layer formation 385, 386
Dipole-layer quenching 527

Direct bandgap 7, 15, 40, 42, 52, 122, 421, 604, 605, 608, 611, 612, 614, 621, 630, 633, 680, 692, 716
Direct lattice 11, 12
Direct transition 604
Direct tunneling 228, 421, 423, 437, 438, 452
Displacement 752
Displacement current 475, 478, 480, 485, 502, 520, 562, 563, 573, 676, 702
Distributed-Bragg reflector 637, 655
Distributed-Bragg reflector laser 636
Distributed-feedback laser 636
DMOS transistor 346, 582, 583
Domain 518, 521, 522, 525, 526, 536
Domain excess velocity 521
Domain excess voltage 522, 534
Domain formation 514, 516, 518
Domain maturity 519
Domain transit time 524
Domain velocity 520
Domain width 522
Donor impurity 21, 22, 80
Donor interface trap 140, 214
Donor level 21
Doping gradient 87
Doping profile 85, 86, 88, 90, 139, 339
Doping superlattice 60
Double-barrier diode 454
Double-diffused MOS transistor 346
Double-drift diode 467, 485, 486, 494
Double-heterojunction bipolar transistor 284, 285
Double-heterojunction laser 621, 627
Double-heterojunction MODFET 410
Double-heterojunction phototransistor 697
Double-heterostructure laser 651
DPS 712
Drain resistance 395, 396, 401
Drain-induced barrier lowering 333
DRAM 349, 350
Drift 28, 480
Drift current 46, 314, 674
Drift mobility 34
Drift region 466, 467, 469, 471, 472, 474, 476, 477, 478, 480, 481, 482, 484, 485, 487, 489, 495, 499, 500, 502, 503, 505
Drift transistor 248
Drift velocity 28, 316, 317, 381, 382, 391, 396, 405, 406, 510, 512, 514, 519
DSP 712
dV/dt triggering 562, 571
Dynamic random-assess memory 349
Dynamic range 671, 701, 712
Dynamic resistance 120

Early voltage 256, 257, 258, 283, 286
ECL 281
Edge effect 112, 555
Edge emitter 618
Edge-emitting laser 655
EEPROM 351, 356
Effective channel length 304, 333, 380
Effective density of states 18, 19
Effective lifetime 565, 566
Effective mass 12, 13, 14, 15, 17, 28, 37, 47, 156, 423, 424, 436, 438, 440, 512, 513, 514, 650
  conductivity 21, 28
  density-of-state 17, 19, 421
  longitudinal 425
  reduced 606
  transverse 425
Effective mobility 317
Effective Richardson constant 47, 128, 156, 161, 162, 170, 174, 440, 499, 532, 540, 714, 734
Effective temperature 36
Effective transverse field 317
Efficiency 469, 482, 484, 485, 486, 487, 488, 494, 497, 504, 525, 526, 527, 529, 531, 643, 725
Einstein coefficient for spontaneous emission 623
Einstein coefficient for stimulated absorption 623
Einstein coefficient for stimulated emission 623
Einstein relation 46, 63, 93, 277, 720
E-k relationship 12, 13, 421, 423, 450, 510
Elastic stiffness 756
Elastic wave 755
Electrically erasable/programmable ROM 351
Electrically programmable ROM 351
Electroluminescence 601, 607, 608
Electromagnetic spectrum 602

Electron affinity 125, 126, 135, 136, 199, 226
Electron gas 404
Electron irradiation 594
Electron temperature 160, 511, 513, 536, 538, 540
Electronegativity 144
Electronic limitation 489
Emission spectrum 606
Emitter efficiency 250, 253, 338
Emitter injection efficiency 252
Emitter resistance 254, 261, 273, 281
Emitter-coupled logic 281
Energy band 12
Energy gap (also see bandgap) 12, 13, 19, 21, 22, 56, 58, 59, 60, 95, 103, 143, 173, 229, 285, 319, 330, 374, 393, 409, 423, 515, 605, 608, 610, 611, 612, 613, 618, 629, 637, 643, 650, 656, 664, 667, 670, 672, 682, 697, 714, 719, 723, 729, 745, 746, 765
Energy relaxation frequency 532
Energy relaxation time 513, 525, 538, 541
Energy-momentum relationship 12, 450
Enhancement mode 296, 326, 327, 375, 390, 392, 404
EOT 340
Epitaxial silicide 181
EPROM 351, 356
Equal-area rule 521
Equivalent circuit 263, 347, 393, 407, 431, 433, 480, 490, 670, 673, 722, 724
Equivalent noise temperature 534
Equivalent oxide thickness 340
Esaki diode 418
Escape probability 718
ESD 120
Eutectic temperature 146, 180
Excess carrier 63, 64, 65, 66
Excess current 419, 428, 429, 430, 431
Excess voltage 535
External quantum efficiency 615, 618, 643, 657
Extinction coefficient 51
Extraction efficiency 615
Extrinsic Debye length 205
Extrinsic photoconductor 664, 667, 670, 671
Extrinsic photoexcitation 667
Extrinsic transconductance 254, 348, 395, 407

Fabry-Perot cavity 634
Fabry-Perot etalon 626
Fabry-Perot resonator 455
Fabry-Perot SOA 657
Face-centered cubic lattice 8, 12
Fall time 575, 576
FAMOS 356
Far-field pattern 643, 646
Fast-recovery diode 122
Fat zero 708
fcc lattice 8, 12
Feedback 445, 483
Fermi distribution function 439
Fermi level 22, 420, 421, 435, 439
Fermi sphere 458
Fermi-Dirac distribution 17, 22, 48, 62, 163, 424, 625, 638, 653
Fermi-Dirac integral 18, 19, 81, 311
Fermi-Dirac statistics 47
FET 294, 296, 374, 396, 418, 539, 541, 542
Fick's law 45
Field emission 165, 608
Field oxide 298
Field-controlled thyristor 591
Field-dependent mobility 307, 308, 317, 335, 382, 383, 406
Field-effect transistor 2, 275, 280, 294, 295, 296, 374, 393, 401, 514, 548
Field-programmable ROM 351
Figure-of-merit 187, 262, 268, 270, 271, 349, 395
Filament formation 489
Fill factor 724, 725, 736
Fixed oxide charge 213, 221, 223, 312, 322
Flash EEPROM 351
Flat-band 185, 199, 201, 202, 205, 216, 223, 312, 375, 378, 498, 500, 713
Flat-band voltage 225, 227, 303, 312, 325, 498, 500, 767
Flicker noise 117, 118, 666
Floating gate 351, 352
Floating-gate avalanche-injection MOS 356
Floating-gate tunnel oxide transistor 356
FLOTOX 356
FM noise 534
$f_{\max}$ 270, 275, 282, 340, 349, 375, 395, 396, 397, 400, 407

F-N tunneling 437, 438
Force 752
Forward blocking 550, 556, 560, 562, 563, 571, 573, 574, 580
Forward recovery time 574
Forward-blocking voltage gain 591, 594
Forward-transmission gain 268
Four-point probe 31
Fowler emission coefficient 682
Fowler theory 176
Fowler-Nordheim tunneling 228, 352, 353, 358, 437, 450, 452
  modified 357, 358
Fractional quantum Hall effect 33
Frame transfer 701
Free-carrier conduction 54
Free-charge transfer model 707
Frenkel-Poole current 359
Frenkel-Poole emission 228, 229
Frenkel-Poole transport 358
Frequency 50, 530, 532, 533, 573, 574, 594, 620, 646, 709
Frequency chirp 649
Frequency deviations (FM) noise 534
Frequency of oscillation 50
Frequency response 676, 756
Fresnel loss 616
Fringing field 707, 711
Fringing-field drift 709
Fringing-field effect 705
f_T 263, 266, 267, 270, 275, 286, 287, 346, 348, 395, 396, 397, 399, 400, 407, 410, 621
Functional device 2, 459, 534
Fusible-link ROM 351

Gain 666, 668, 669, 671, 683, 685, 687, 689, 693, 694, 696, 697, 713, 714
Gain saturation 639
Gain-guided 635
Gas sensor 765, 766
Gate capacitance 380, 396
Gate current 336
Gate dielectric 297, 340
Gate length 376, 396, 399
Gate oxide 298
Gate resistance 395, 399, 400
Gate stack 340
Gate turn-off thyristor 574
Gate-assisted turn-off 573
Gate-assisted turn-off thyristor 576
Gate-channel capacitance 395
Gated diode 112, 338
Gate-drain capacitance 396
Gate-induced drain leakage 114, 338
Gate-source capacitance 385, 392
Gauge factor 750, 751
Gauss' law 62, 141, 202, 223, 301, 302, 353
Gaussian mode 646
Gaussmeter 758
Generalized scaling 330
Generation 40, 63, 96, 97, 445, 567
Generation current 97, 98, 122
Generation lifetime 44, 96
Generation rate 38, 45, 63, 474, 674
Generation-recombination center 573
Generation-recombination noise 667, 670
Geometric correction factor 760
Geometric effect 750, 760
Geometric magnetoresistance effect 760, 761
GIDL 114, 338
Glow-discharge 732
g_m 254, 262, 263, 264, 275, 309, 312, 317, 337, 346, 348, 349, 385, 391, 392, 395, 396, 399, 400, 407, 541, 542
Graded HBT 284
Graded-base 285
Graded-base bipolar transistor 284, 286
Graded-channel FET 399
Gradient voltage 88
Gradual-channel approximation 303, 316, 328, 377
Grain boundary 765, 766
Grating 716
Ground-state degeneracy 22
Group velocity 12, 14
GTO 574, 576
Guard ring 172, 182, 689, 691
Gummel number 249, 253, 257, 258, 259
Gunn diode 418, 511
Gunn effect 511
Gunn oscillation 511, 516, 524, 525

Hall angle 762
Hall coefficient 34
Hall effect 30, 33, 758, 759
Hall factor 34, 759

Hall field 33, 762
Hall generator 758
Hall mobility 30, 34
Hall plate 758, 761, 762, 763
Hall voltage 34, 759, 760, 761, 763
Haynes-Shockley experiment 66
HBT 127, 267, 282, 285
Heat sink 279, 375, 472, 488, 489, 530, 548, 568, 651
HEMT (see also MODFET) 401
  P-HEMT 408
Hermite polynomial 646
Hermite-Gaussian distribution 646
Hermite-Gaussian function 646
HET 450
Heteroepitaxy 57, 454
Heterointerface 374, 375, 401, 402, 403, 406, 407, 408, 410, 510
Heterojunction 1, 56, 79, 124, 189, 375, 401, 409, 452, 538, 612, 630
  anisotype 124
  broken-gap 58
  isotype 124, 127
  staggered 58
  straddling 58
Heterojunction APD 693
Heterojunction avalanche photodiode 692
Heterojunction bipolar transistor 127, 267, 282
Heterojunction FET 295, 374
Heterojunction field-effect transistor 401
Heterojunction insulated-gate FET 374
Heterojunction laser 621, 626
Heterojunction photodiode 671, 680
Heterojunction phototransistor 696, 697
Heterojunction THETA 451
Heterostructure 401, 536
Heterostructure APD 692
Heterostructure laser 628, 629, 640
Hexagonal close-packed lattice 8
HFET 295, 374, 401
HIGFET 374, 408
High-electron-mobility transistor 401
High-energy particle 725
High-field property 35
High-injection 96, 99, 123, 251, 259
High-K dielectric 340
High-level injection 43, 44, 246, 253, 254, 277, 472, 565, 615
High-low profile 320, 321
High-threshold state 352
Hi-lo diode 467
Holding current 550, 573, 574
Holding voltage 447, 448, 550, 570, 571
Homostructure laser 628
Hook collector 549
Hooke's law 50, 752
Hot carrier 335, 450, 453, 538, 539, 603
Hot spot 278
Hot-carrier injection 352, 354
Hot-electron injection 351, 375
Hot-electron scattering 450
Hot-electron spectroscopy 459
Hot-electron transistor 287, 450
Hot-electron trapping 375
Hydrogen-atom model 21
Hydrostatic pressure 650
Hyper-abrupt 121, 122

Ideality factor 98, 119, 152, 164, 170, 395, 440, 725, 734
IDT 754
IGBT 2, 548, 582
IGFET 295, 582
IGR 582
IGT 582
IIL, I^2L 281, 282
Image charge 147
Image force 147
Image-force dielectric constant 149, 178
Image-force lowering 136, 146, 151, 152, 156, 170, 172, 178, 228, 715
Image-force permittivity 149
Impact 753
Impact ionization 37, 40, 63, 104, 105, 335, 337, 352, 388, 466, 487, 549, 608, 671
Impact-ionization avalanche transit-time diode 466
IMPATT diode 2, 105, 184, 418, 466, 530, 532, 534, 690
Impedance 433
Impurity 16
Impurity gradient 110
Impurity scattering 34, 35, 37, 374, 401, 402, 403, 421, 427, 456

Impurity-band conduction 429
Incandescence 601
Incoherent 623
Index of ballisticity 312
Index-guided 635, 636, 645
Indirect bandgap 15, 42, 427, 604, 605, 608, 611, 612, 614, 629, 630, 633, 716, 732
Indirect transition 604
Indirect tunneling 421, 427
  phonon assisted 427
Induced base 452
Inductance 431
Inductor 483
Inelastic phonon scattering 456
Infrared 611
Injection delay 476, 482
Injection efficiency 170, 258, 282, 445, 553, 561, 562, 695, 696, 697
Injection electroluminescence 607
Injection laser 621
Injection phase delay 474, 485
Injection ratio 166, 168
Injection velocity 311
Injection-modulation mode 763
Input capacitance 396
Input reflection coefficient 268
Input resistance 395
Insulated-gate bipolar transistor 2, 548, 582
Insulated-gate FET 295, 582
Insulated-gate rectifier 582
Insulated-gate transistor 582
Integrated-injection logic 281
Integration time 701
Interband recombination transition 652
Interband transition 603, 605, 608, 633, 656
Interband tunnel diode 418
Interband tunneling 422
Interdigital transducer 754
Interdigitated 569, 576, 713
Interface scattering 398
Interface state 139, 140, 213
Interface trap 118, 170, 198, 213, 216, 217, 303, 337, 375, 442, 443, 633, 707, 708, 710, 711
  acceptor 140, 141, 214
  donor 140, 214
Interface-trap density 141, 215, 216, 220, 221, 315, 316, 337, 359
Interface-trap lifetime 215
Interface-trapped charge 221
Interline transfer 701
Internal photoemission 681, 682
Internal quantum efficiency 614, 615, 621, 699, 713
Intersubband excitation 717
Intersubband transition 633, 656
Intervalley scattering 28
Intervalley scattering time 525
Intraband transition 603, 608
Intravalley scattering 28
Intrinsic avalanche response time 488
Intrinsic concentration 19, 279, 283, 286, 746
Intrinsic excitation 607, 608
Intrinsic photoconductor 667, 669
Intrinsic photoexcitation 667
Intrinsic transconductance 254
Inversion 200, 298, 299, 301, 306, 312
Inversion charge 298, 303, 311
Inversion layer 296, 316, 445, 447
Inverted MODFET 410
Inverter 349
Ionic conduction 228
Ionization coefficient 337, 409, 471, 485, 491, 549, 683, 684, 685, 686, 687, 690
Ionization energy 21
Ionization integral 105, 106
Ionization integrand 467, 469
Ionization rate 37, 39, 40, 467, 471, 472, 478, 486, 487, 488, 492, 493
Ion-sensitive field-effect transistor 766
ISFET 766
Isoelectronic center 612, 613, 616
Isoelectronic impurity 608, 612
Isoelectronic trap 612
Isotype heterojunction 124, 127

JFET 2, 295, 296, 328, 374, 375, 548, 587
Johnson noise 117, 666
Joint density of states 606
Joint dispersion relation 606
Junction curvature 172
Junction FET 295, 374
Junction laser 621
Junction transistor 243

$k \cdot p$ method 13

Kink effect 338, 344
Kirchhoff's law 251, 540
Kirk effect 259, 261, 266, 276, 286
k-selection rule 604, 605
k-space 510

Lambertian emission pattern 617
Landau level 650
Large-optical-cavity heterostructure laser 634
Large-signal operation 482
LASCR 580, 582
Laser 3, 124, 601, 609, 621
  DBR 636
  DFB 636, 637
  DH 634, 643, 644, 651
  distributed-Bragg reflector 636
  distributed-feedback 636
  double-heterojunction 621, 627
  double-heterostructure 634, 651
  edge-emitting 655
  heterojunction 621, 626
  heterostructure 628, 629, 640
  homostructure 628, 634
  large-optical-cavity heterostructure 634
  LOC heterostructure 634
  quantum cascade 633, 656
  quantum-dot 654
  quantum-well 651, 654
  quantum-wire 654
  SCH 634, 636, 653
  separate-confinement heterostructure 634, 653
  single-heterostructure 634
  stripe geometry 635, 643, 645, 646
  superlattice 653
  vertical-cavity surface-emitting 655
Laser degradation 651
Laser diode 621
Latch-up 342, 344
Lateral insulated-gate transistor 582
Laterally diffused MOS transistor 347
Lattice conduction 54
Lattice constant 8, 12, 56, 680
Lattice mismatch 57
Lattice temperature 160, 513, 514, 515, 536, 538, 540
Lattice vibration 51
LDD 340
LDMOS 347
LED 3, 124, 601, 604, 608, 641
  white-light 619
Lifetime 43, 563, 565, 620, 622, 675
  ambipolar 566
  radiative recombination 610
Light amplification by stimulated emission of radiation 601, 621
Light-activated switch 580
Light-activated thyristor 580
Light-emitting diode 3, 124, 601, 608
Light-escape cone 616
Lightly doped drain 340
Light-triggered switch 447, 448
LIGT 582
Limited-space-charge accumulation mode 529
Linear region 376, 391
Linearity 392, 696
Linearly graded 86, 88, 96, 107, 110, 112
LOC heterostructure laser 634
Lo-hi-lo diode 467, 532
Long-channel behavior 379
Long-channel MOSFET 294, 319, 329
Longitudinal effective mass 425
Longitudinal elastic constant 28
Longitudinal field 303, 307, 316, 317, 328
Longitudinal mode 626, 642
Longitudinal optical-phonon energy 652
Longitudinal piezoresistive coefficient 751
Longitudinal wave 755
Lorentz force 33, 761
Loudspeaker 753
Low-field mobility 49, 307, 312, 316, 317, 318, 385, 396, 487, 514
Low-high profile 320, 323
Low-injection 63, 68, 90, 93, 99, 500, 726
Low-level injection 42, 43, 63, 92, 614, 615
Low-threshold state 352
LSA mode 529
Luminescence 602
Luminous efficiency 602, 618, 619
Luminous flux 619

MAGFET 763
Magnetic field 629, 650, 743, 760, 762, 763
Magnetic sensor 3, 743, 758
Magnetic-field sensor 760

Magnetic-field-sensitive field-effect transistor 763
Magnetodiode 762
Magnetometer 758
Magnetoresistance effect 35, 760
  geometric 760, 761
  physical 760
Magnetoresistor 760, 761
Magnetotransistor 762
Magnistor 762
Maser 621
Mask-programmed ROM 351
Mass-action law 21, 22
Matthiessen rule 28
Maximum available power gain 268, 270
Maximum field 82, 87, 107, 110, 137, 151, 152, 470, 471, 472, 690
Maximum frequency of oscillation ($f_{\max}$) 270, 349, 395, 396, 407
Maximum power 723
Maxwell equation 62, 628
Maxwellian 540
Maxwellian distribution 157, 161
MBE 285, 431, 451, 454
Mean free path 28, 37, 46, 161, 309, 386, 452
Mean free time 28, 37
Measurand 743, 744
Mechanical sensor 3, 743, 744, 750
Merged-transistor logic 281
MESFET 2, 104, 135, 295, 296, 328, 349, 374, 375, 548, 587
Metal oxide 745
Metal-base transistor 287
Metal-insulator-metal tunnel diode 448
Metal-insulator-metal-insulator-metal structure 450
Metal-insulator-metal-semiconductor structure 450
Metal-insulator-semiconductor capacitor 1, 197, 699
Metal-insulator-semiconductor FET 295
Metal-insulator-semiconductor solar cell 734
Metal-insulator-semiconductor structure 437
Metal-insulator-semiconductor tunnel diode 169
Metal-nitride-oxide-silicon transistor 357
Metal-organic chemical vapor deposition 431
Metal-oxide semiconductor 765
Metal-oxide sensor 765
Metal-oxide thermistor 745
Metal-oxide-metal-oxide-metal structure 450
Metal-oxide-metal-semiconductor structure 450
Metal-oxide-nitride-oxide-silicon transistor 360
Metal-oxide-semiconductor field-effect transistor 293
Metal-oxide-semiconductor structure 1
Metal-oxide-silicon 197
Metal-oxide-silicon capacitor 213
Metal-semiconductor APD 689
Metal-semiconductor barrier 608
Metal-semiconductor contact 1, 120, 128, 129, 134, 467, 497
Metal-semiconductor diode 122
Metal-semiconductor FET 295, 374
Metal-semiconductor junction 374, 377
Metal-semiconductor photodiode 664, 671, 679, 680, 682
Metal-semiconductor-metal photodetector 712
Metal-semiconductor-metal structure 497
Metamorphic MODFET 409
Microphone 753
Microplasmas 689
Microwave 349, 466, 467, 485, 495, 511, 524, 534, 671, 683
Microwave amplification by stimulated emission of radiation 621
Microwave detection 435
Microwave performance 375, 393, 407
Miller indices 8
MIM diode 448
MIM tunnel diode 448
MIMIM 450
MIMS 450
Miniband 657
Minimum noise figure 398, 407
Minimum noise measure 491, 492
Minority-carrier diffusion length 725
Minority-carrier injection 166, 272
Minority-carrier lifetime 44, 65, 68, 119, 122, 250, 273, 445, 573, 581, 585, 594, 646, 695, 709, 725
Minority-carrier storage 296, 435, 453, 486, 488

Minority-carrier storage time 168
Minority-charge storage 459
MI*p-n* 450
MIS 733
MIS capacitor 1, 197, 437, 699
MIS solar cell 734
MIS structure 437
MIS switch diode 444
MIS thyristor 447
MIS tunnel devices 437
MIS tunnel diode 169, 437, 439, 442, 444, 768
Misawa diode 467, 484
MISFET 295
MISS 444
Mixer 182, 407
Mixing 435
M-MODFET 409
MNOS Transistor 357
Mobile ionic charge 213, 223
Mobile oxide charge 303
Mobility 28, 35, 296, 307, 312, 316, 320, 337, 339, 379, 397, 401, 406, 510, 512, 514, 536, 549, 590, 669, 710, 711, 718, 746
  conductivity 30
  constant 303, 306, 307, 308, 312, 326, 379, 381, 383, 384, 391, 392, 405
  differential 514, 519
  drift 34
  effective 317
  field-dependent 307, 308, 317, 328, 335, 382, 383, 406
  Hall 30, 34
  low-field 49, 307, 312, 316, 318, 385, 396, 487, 514
  negative differential 37, 511, 514, 515, 516, 519, 530, 534
  positive differential 530
Mobility modulation 539, 542
MOCVD 285, 431, 451, 454
MODFET 2, 374, 398, 401, 539, 548
  double-heterojunction 410
  inverted 410
  metamorphic 409
  M-MODFET 409
  P-MODFET 408, 409
  pseudomorphic 408
  quantum-well 410
  superlattice 410
Modified Fowler-Nordheim tunneling 357, 358
Modified Read diode 467, 469
Modulating sensor 744
Modulation doping 374, 401, 403, 409, 410, 452
Modulation factor 485
Modulation frequency 670, 677, 678
Modulation index 670
Modulation-doped channel 401
Modulation-doped FET 374
Modulation-doped field-effect transistor 401
Modulation-doped superlattice 401
Molecular-beam epitaxy (MBE) 431
Momentum 428, 439, 457, 604, 612
MOMOM 450
MOMS 450
MONOS transistor 360
Monte Carlo 515
MO*p-n* 450
MOS 197
MOS capacitor 1, 213, 303, 768
MOS structure 1
MOSFET 1, 2, 104, 114, 197, 221, 275, 293, 374, 380, 450, 539, 548, 582, 584, 586, 763, 766, 767, 768
  buried-channel 326, 327
  long-channel 329
  power 346
  three-dimensional 345
MOSFET scaling 329
Mott barrier 134, 185
Mott-Gurney law 49
MSM photodetector 712
MSM structure 497, 499
Multiplication 37, 79, 96, 102, 104, 120, 231, 232, 250, 335, 346, 390, 447, 467, 471, 474, 479, 482, 492, 551, 552, 557, 558, 663, 671, 683, 685, 688, 690, 691
Multiplication factor 105, 257, 683
Mushroom-gate 400

Nanostructure 56
n-channel 296, 297, 298
NDR 454, 458, 459, 511, 514, 524, 536, 540
Negative differential mobility 37, 511, 514, 515, 516, 519, 530, 534

Negative differential resistance 2, 103, 418, 421, 429, 445, 454, 473, 474, 510, 511, 514, 516, 524, 536, 540, 548, 578, 637, 746
Negative resistance 382, 386, 437, 443, 444, 456, 466, 476, 477, 482, 483, 491, 515, 529, 537
Negative temperature coefficient 104, 120, 296, 745
Negative-resistance field-effect transistor 538
NEP 667, 670, 674, 689
NERFET 538
Neutral level 140, 141, 144, 214
Neutron 553
Neutron irradiation 44, 553
Neutron transmutation 553
NMOS logic 349
Noise 117, 340, 375, 398, 399, 407, 467, 483, 489, 497, 532, 534, 590, 663, 666, 671, 672, 679, 683, 686, 687, 690, 691, 692, 693, 694, 697, 700, 703, 709
  1/*f* 118, 271, 667, 763
  AM 534
  amplitude deviations (AM) 534
  flicker 117, 118, 666
  FM 534
  frequency deviations (FM) 534
  generation-recombination 667, 670
  Johnson 117, 666
  shot 117, 118, 666, 670, 672, 674, 685, 687, 690
  thermal 117, 118, 271, 666, 670, 673, 688
  white 118
Noise constant 434
Noise current 118, 490, 491
Noise equivalent power 689, 696
Noise factor 685, 686, 690, 693
Noise figure 271, 275, 398, 400, 407, 434, 490, 491
  minimum 398, 407
Noise measure 490, 491, 492, 493, 497, 504
  minimum 491, 492
Noise source 398
Noise voltage 491
Noise-equivalent power 667, 674
Nonlinear region 376, 380
Nonradiative lifetime 614, 615, 621
Nonradiative recombination center 651
Nonradiative recombination rate 614
Nonradiative transition 603, 614
Nonvolatile memory 1, 2, 350, 366
Nonvolatile RAM 351
Normal mode 245, 247, 251, 277
Normally-off 296, 326, 392, 404, 593, 594
Normally-on 296, 326, 376, 392, 404, 586, 587, 593
NTC 745, 746

Occupancy 425
Occupation probability 163
Offset voltage 285, 760
Off-state 444, 445, 548, 549, 550, 563, 571, 574, 577
Ohmic contact 1, 135, 153, 162, 164, 166, 185, 187, 188, 189, 376, 403, 431, 487, 530, 531, 667, 726, 728
One-sided abrupt junction 80, 83, 85, 106
ONO 360
On-resistance 585, 586, 590, 593
On-state 271, 273, 444, 445, 548, 549, 550, 560, 562, 563, 564, 565, 571, 572, 575, 577, 581, 583, 585
Open-base 257
Open-circuit voltage 170, 722, 723, 731, 734, 735, 736
Optical cavity 636, 655
Optical concentration 735
Optical confinement 635, 640, 653
Optical efficiency 615, 616, 617
Optical excitation 63
Optical gain 623, 626, 638, 639, 641, 655, 718
Optical generation 447
Optical mode 50
Optical phonon 28, 36
Optical pumping 622
Optical radiation 602, 607
Optical resonator 626, 627, 642
Optical sensor 743, 744
Optical-fiber communication 3, 609, 610, 622, 630, 633, 641, 646, 649, 657, 663, 664
Optical-phonon energy 36
Optical-phonon scattering 156, 161
Opto-isolator 609
Orthodox theory 365
Orthogonalized plane-wave method 13
Oscillation 467

Oscillator 407, 418, 448, 483, 484, 492, 497, 511
Output capacitance 395
Output reflection coefficient 268
Output resistance 395
Overshoot 386
Oxide charge 337
Oxide trapped charge 213, 223, 224, 225

Parallel injection 701
Parasitic input capacitance 395, 396
Passive pixel sensor 712
Pauling's electronegativity 144
p-channel 296, 297
Peak current 425, 430, 431, 433, 459
Peak velocity 515
Peak voltage 419, 425, 427, 431, 433, 458
Peak-to-valley ratio 431, 456, 525, 538, 540, 541
Pedestal collector 275
Peltier effect 748
Peltier EMF 748
Pentode-like 591
Permeable-base transistor 587
PET 294
Phase delay 476
P-HEMT 408
Phonon 51, 611
Phonon conduction 54
Phonon mean free path 55
Phonon scattering 29, 34, 35, 401, 421, 427, 456, 746
Phonon spectra 50
Phonon velocity 55
Phonon-assisted indirect tunneling 427
Phonon-assisted tunneling 421, 434, 456
Phonon-phonon scattering 56
Phosphor 619
Photoconductive effect 44, 752
Photoconductive gain 718
Photoconductivity 713, 717
Photoconductor 667, 718
  extrinsic 664, 667, 670, 671
  intrinsic 667
Photocurrent 176, 581, 582, 666, 669, 671, 674, 676, 683, 686, 691, 696, 699, 700, 712, 718, 722, 723, 725, 726, 728, 729, 730, 733
Photo-Darlington 697
Photodetector 3, 135, 601, 604, 622, 633, 657, 663, 667, 744
  metal-semiconductor-metal 712
  MSM 712
  quantum-well infrared 716
Photodiode 633, 669, 671, 712, 720
  avalanche 633, 671, 681, 683, 694
  heterojunction 671, 680
  metal-semiconductor 664, 671, 679, 680, 682
  p-i-n 671, 674, 678, 681, 687, 715
  p-n 671, 674, 679
  point-contact 682
  Schottky-barrier 671, 682, 713, 715
Photoelectric effect 664, 666
Photoelectric measurement 176, 178
Photoelectromagnetic effect 44
Photoemission spectroscopy 144
Photoexcitation 664, 667
Photoluminescence 607
Photon detector 664
Photon energy 625
Photon flux 666, 674, 700, 718, 720
Photon lifetime 648
Photoresponse 178
Phototransistor 594, 694, 713
  Darlington 697
  double-heterojunction 697
  heterojunction 696, 697
Photovoltaic 601, 719, 744
Physical magnetoresistance effect 760
Piezoelectric charge constant 753
Piezoelectric crystal 753
Piezoelectric effect 754, 755, 760
Piezoelectric material 753, 754, 756
Piezoelectric microphone 753
Piezoelectric polarization 409
Piezoelectric speaker 753
Piezoelectric strain gauge 753
Piezoelectric transducer 753
Piezoelectricity 753
Piezoresistance 752
Piezoresistive effect 750, 751
Piezoresistive strain gauge 753
Piezoresistivity 760
p-i-n diode 123, 467, 469, 473, 474, 557, 584, 591, 593, 762

p-i-n photodiode 671, 674, 678, 681, 687, 715
Pinch-off 304, 306, 307, 316, 379, 380, 381, 391, 405, 406, 587, 589, 590, 593
Pinch-off potential 378, 390
Planar process 182, 431
Planar technology 689
Planar-doped-barrier transistor 287
P-MODFET 408, 409
p-n junction 1, 79, 123, 181, 182, 295, 296, 315, 374, 375, 376, 418, 436, 444, 445, 466, 467, 471, 497, 550, 564, 582, 587, 607, 608, 624, 654, 691, 712, 725, 733, 748
p-n photodiode 671, 674, 679
p-n-p-n 549
Point contact 134, 181
Point-contact photodiode 682
Point-contact rectifier 181
Point-contact transistor 243
Poisson equation 49, 62, 81, 83, 86, 106, 112, 121, 126, 136, 201, 203, 207, 233, 259, 265, 321, 377, 473, 499, 518, 519, 555
Poisson's ratio 751
Polar optical scattering 515
Polarization 755
Polarization correction factor 718
Polarization selection rule 716
Polar-optical-phonon scattering 28
Poly-emitter 273, 274
Population inversion 624, 625
Positive differential mobility 530
Positive differential resistance 473
Positive temperature coefficient 104, 120, 745, 748
Potential-effect transistor 2, 294, 541, 548
Power 375, 393, 482, 483, 484, 485, 486, 493, 497, 504, 532, 533, 548, 643, 723
Power amplifier 511, 548, 590
Power conversion efficiency 720
Power devices 548
Power efficiency 618
Power gain 268, 270, 395, 490
Power MOSFET 346
Power transistor 275, 279
Power-frequency limitation 397, 483, 488
Poynting vector 629
PPS 712
Pressure 431, 515, 650, 752, 757
Pressure transducer 752
Primary photocurrent 669
Primitive basis vector 10
Primitive cell 8, 10, 11, 12
Programmable ROM 351
PROM 351
Proton bombardment 635
Proton irradiation 594
Pseudomorphic MODFET 408
Pseudopotential method 13
PTC 745, 746
Pulse bond 431
Punch-through 111, 256, 320, 328, 331, 333, 334, 335, 339, 344, 346, 447, 551, 552, 553, 587, 594, 713

Quantized energy 59
Quantized level 611, 652
Quantized state 455
Quantum cascade laser 633, 656
Quantum dot 60, 61, 361
Quantum efficiency 611, 612, 651, 653, 664, 666, 668, 669, 671, 672, 674, 675, 676, 678, 680, 682, 683, 689, 690, 692, 693, 713, 715, 718, 720, 729
Quantum well 58, 61, 124, 438, 454, 456, 611, 656, 716
Quantum wire 60, 61
Quantum-dot laser 654
Quantum-mechanical tunneling 417
Quantum-well infrared photodetector 716
Quantum-well laser 651, 654
Quantum-well MODFET 410
Quantum-wire laser 654
Quasi-constant-voltage scaling 330
Quasi-Fermi level 91, 298, 300, 624
Quasi-field 286
Quasi-neutral region 169
Quasi-saturation 276, 277, 286
Quenched dipole-layer mode 526
QWIP 716

Radiative 603
Radiative lifetime 614, 615, 621
Radiative recombination 607, 608, 610, 611, 614
Radiative recombination center 612
Radiative recombination lifetime 610

Radiative recombination rate 614
Radiative transition 604, 605, 607, 630
Radioluminescence 607
Radius of curvature 112
Raised source/drain 343
RAM 351, 407
Raman scattering 51
Random access 712
Random-access memory 351
RC time constant 527
Reach-through 497, 499, 500, 690
Reach-through diode 497
Reactive cutoff frequency 433
Read diode 467, 469, 471, 472, 477, 482, 484
  modified 467, 469
Read-write memory 351
Real-space transfer 510, 536
Real-space-transfer 2
Real-space-transfer diode 514, 536
Real-space-transfer transistor 538
Recessed-channel 399, 400
Recessed-gate 399
Reciprocal lattice 10, 11, 12
Recombination 40, 63, 96, 97, 99, 123, 153, 154, 249, 250, 443, 445, 567, 568, 572, 584, 603, 607, 610, 672
Recombination center 122, 273, 573, 605, 615, 622
Recombination coefficient 40, 614, 615
Recombination current 97, 98, 119, 245, 250, 253, 273, 562, 567, 584, 593, 725, 731, 748, 762
Recombination lifetime 44, 732
Recombination rate 43, 63, 93, 565, 668
Recombination trap 565
Recombination velocity 159, 160, 161, 731
Rectangular barrier 47, 439
Rectifier 119
Reduced effective mass 606
Reference resistance 745
Reference temperature 749
Reflection 156, 161, 451, 452, 616, 674, 718
Reflection coefficient 53, 616
Reflectivity 731
Refraction 615, 616
Refractive index 51, 616, 622, 627, 628, 629, 634, 636, 637, 640, 649, 650
Regenerative feedback 445
Relative eye sensitivity 602, 619
Relaxation oscillation frequency 649
Reset transistor 712
Resistance 431
  channel 395, 397
  differential 185, 538
  differential negative 37
  drain 396
  gate 395, 399, 400
  input 395
  negative 437, 443, 444, 456, 466, 476, 477, 482, 483, 491, 515, 529, 537
  negative differential 2, 103, 418, 421, 429, 431, 445, 454, 473, 474, 510, 511, 514, 516, 524, 536, 540, 548, 578, 637, 746
  output 395
  positive differential 473
  source 395, 396, 399, 407
  space-charge 436, 474
  spreading 431
  thermal 279, 488, 489
  tunneling 361
Resistance temperature detector 744, 748
Resistive cutoff frequency 433
Resistivity 30, 431, 553
Resistor
  thermally sensitive 744
  voltage-controlled 375, 376
Resonant circuit 525, 527, 529
Resonant frequency 480, 486, 491, 626
Resonant SOA 657
Resonant tunneling 455, 458, 656
Resonant-tunneling bipolar transistor 459
Resonant-tunneling current 455
Resonant-tunneling diode 2, 418, 454
Resonant-tunneling hot-electron transistor 459
Responsivity 666, 671
RESURF 347, 390
Retention time 356
Retrograde profile 323, 339
Reverse blocking 550, 551, 552, 553, 559
Reverse-transmission gain 268
Ribbon growth 730
Richardson constant 156
  effective 156, 162
Ridley-Watkins-Hilsum effect 511
Rock-salt lattice 8

ROM 351
RST diode 536, 538
RST transistor 538
RTD 748

Safe operating area 279
Salicide 341, 346
Satellite valley 511, 514, 537
Saturation current 311, 376, 385, 387, 406, 723
Saturation mode 247, 250, 255, 257
Saturation region 303, 306, 376, 380, 384, 391, 395, 397
Saturation velocity 36, 37, 63, 261, 307, 309, 310, 312, 318, 379, 382, 385, 386, 387, 396, 397, 472, 473, 474, 477, 478, 482, 484, 487, 496, 499, 500, 521, 542, 549, 685, 690
Saturation voltage 376
SAW 754
Scaling 328
Scaling factor 330
Scaling limit 339
Scattering 40, 309, 386, 421, 452, 532, 533
  carrier-carrier 566, 568
  Coulomb 28
  hot-electron 450
  impurity 34, 35, 37, 374, 401, 402, 403, 421, 427, 456
  interface 398
  intervalley 28
  intravalley 28
  optical-phonon 156, 161
  phonon 29, 34, 35, 401, 421, 427, 456, 746
  phonon-phonon 56
  polar optical 515
  polar-optical-phonon 28
  Raman 51
  surface 398
Scattering parameter 267
Scattering time 525
SCCD 699, 706, 710
SCH laser 634, 636, 653
Schottky barrier 134, 142, 151, 295, 375, 377, 436, 440, 531, 532, 712, 733
Schottky diode 122, 154, 272, 682
Schottky effect 146
Schottky emission 228
Schottky junction 374, 376
Schottky source/drain 341, 342, 343
Schottky-barrier APD 691, 692
Schottky-barrier clamp 272
Schottky-barrier diode 46, 437, 682, 733, 768
Schottky-barrier lowering 146, 152, 171
Schottky-barrier photodiode 671, 682, 713, 715
Schottky-barrier solar cell 170
Schottky-barrier source/drain 341
Schrödinger equation 12, 13, 47, 58, 455
SCL current 589
SCR 550, 582, 583, 585
Screen printing 731
SDHT 401
Second breakdown 278, 279, 280, 296
Seebeck effect 748
Seebeck voltage 748, 749
Selectively doped heterojunction transistor 401
Self-aligned silicide 341
Self-generating 753
Self-generating sensor 744
Self-heating 746, 752
Self-induced drift 705, 706, 707
Self-ordering 654
Semiconductor laser amplifier 657
Semiconductor optical amplifier 657
Semiconductor sensor 743, 744
Semiconductor-controlled rectifier 550
Sensitivity 668, 712
Sensor 3, 743
  acoustic 754
  capacitive 757
  chemical 3, 743, 765, 768
  gas 765, 766
  magnetic 3, 743, 758
  magnetic-field 760
  mechanical 3, 743, 744, 750
  modulating 744
  optical 743, 744
  self-generating 744
  semiconductor 744
  temperature 749
  thermal 3, 743, 744
Separate absorption and multiplication 692
Separate-confinement heterostructure laser 634, 653

Separation by implantation of oxygen 344
Sequential injection 701
Sequential tunneling 456
Series resistance 79, 96, 100, 182, 184, 185, 246, 271, 330, 339, 340, 343, 347, 431, 486, 487, 673, 679, 682, 683, 715, 724, 725, 726
  drain 395, 401
  source 395, 401
SET 360
Shallow impurity 21, 29
Shallow-level impurity 176
SHBT 285
Shear wave 755
Sheet resistance 31, 191
Shift register 698, 701, 703, 704
Shock 753
Shockley diode 444, 550, 569, 571, 578
Shockley equation 90, 95, 96, 97, 118
Shockley-Read-Hall recombination 249, 250
Shockley-Read-Hall statistics 42
Short-channel effect 294, 304, 328, 329, 331, 333, 339, 340, 341, 343, 348
Short-circuit current 45, 723, 731, 732, 734, 735
Shot noise 117, 118, 666, 670, 672, 674, 685, 687, 690
Signal processing 757
Signal-to-noise ratio 666, 667, 670, 671, 673, 687, 688, 689, 690, 709, 712
Silicide 146, 178, 180, 181, 298, 341, 342, 682
Silicon isotope 553
Silicon-on-insulator 343
Silicon-on-nothing 344
Silicon-on-oxide 344
Silicon-on-sapphire 344, 762
Silicon-on-zirconia 344
Silicon-oxide-nitride-oxide-silicon transistor 360
SIMOX 344
Single mode 642
Single-drift diode 486, 494
Single-electron box 362, 363, 366
Single-electron island 360, 364, 366
Single-electron transistor 360
SIT 586
SIThy 591
Skin effect 486, 487
Small-signal analysis 477
Snap-back 548
Snapback diode 122
Snell's law 615, 616
SOA 279, 657
SOI 338, 343, 344
Solar cell 3, 135, 170, 601, 604, 622, 667, 719
  thin-film 732
Solar panel 719
Solar radiation 720, 722
Solar spectrum 720, 721, 732
SONOS Transistor 360
SOS 344, 762
Sound wave 753
Source resistance 395, 396, 399, 401, 407
SOZ 344
Space charge 137, 141, 200, 481
Space-charge capacitance 264, 265
Space-charge density 205
Space-charge effect 48, 49, 141, 229, 467, 472, 473, 474, 486, 499, 500, 541, 670, 683
Space-charge resistance 436, 474
Space-charge-limited current 49, 229, 335, 473, 586, 732
Space-charge-limited transport 497
S-parameter 267
Specific contact resistance 187, 188
Specific heat 55
Spectral response 729
Spectrometer 287
Spectrum 606
Speed 663, 671, 672, 678, 692, 696, 715
Speed index 433
Spherical *p-n* junction 112
Spin-orbit interaction 13
Spin-orbit split 693
Spontaneous emission 606, 622, 623, 625, 634, 639, 641, 642, 648
Spontaneous polarization 409
Spontaneous recombination rate 647
Spreading resistance 182, 185, 189, 431
Spreading velocity 568, 569
SRAM 349, 350, 448
Stabilizing resistor 277
Staggered heterojunction 58
Standing wave 626
Static induction 587, 589, 594

Static random-access memory 349
Static-induction current 594
Static-induction device 548
Static-induction transistor 586, 591
Static-inductor thyristor 591
Step-recovery diode 122
Stevenson-Keyes method 65
Stimulated emission 621, 622, 623, 625, 626, 634, 638, 639, 643, 646, 647, 653
Stimulated-emission recombination rate 647
Storage phase 115
Storage time 272, 273, 575, 576
Storage time delay 575
Straddling heterojunction 58
Strain 651, 750, 751, 752, 753
Strain gauge 750
  piezoelectric 753
  piezoresistive 753
Strained layer 57
Stress 750, 752, 753
Stripe-geometry laser 645, 646
Strong inversion 201, 202, 205, 207, 208, 209, 301, 321
S-type negative differential resistance 548
Subband 454
Substrate bias 313
Substrate current 337
Subthreshold 314, 315, 319, 334
Subthreshold current 314, 315, 327
Subthreshold slope 315, 316
Subthreshold swing 315, 319, 323, 327, 335, 337, 344
Sun 720
Superlattice 58, 60, 124, 401, 410, 454, 656, 717
  doping 60
Superlattice laser 653
Superlattice MODFET 410
Surface channel 296, 326, 337, 375
Surface electric field 301
Surface emitter 618
Surface field-effect transistor 293
Surface generation velocity 709
Surface potential 148, 200, 202, 205, 216, 217, 218, 221, 233, 300, 302, 316, 321, 330, 388, 445, 699, 700, 706, 707, 708, 709, 767
Surface recombination 67, 699, 727, 728, 730
Surface recombination velocity 67, 68, 167, 709, 727, 730
Surface scattering 328, 398
Surface trap 253, 388, 667
Surface-acoustic-wave transducer 754
Surface-channel CCD 699, 710
Sweep-out 669
Switch 548, 562
  diode ac 577
  light-activated 580
  light-triggered 448
  triode ac 577
Switching 271, 447, 448, 548, 549, 557, 560, 564, 565, 570
Switching current 550, 581
Switching speed 433
Switching time 119
Switching voltage 444, 447, 448, 550, 571, 578
Symbols 775

Taylor's expansion 379, 380
TE wave 629, 643
TED 510, 511
TEGFET 401
Temperature coefficient 398
Temperature coefficient of resistance 745
Temperature dependence 318
Temperature effect 472
Temperature gradient 54
Temperature sensor 749
Tetrahedral phase 8
Textured cell 731
TFT 338, 343, 345
T-gate 400, 403
Thermal conductivity 54, 55, 56, 489, 496, 549, 582
Thermal detector 664
Thermal diffusion 705
Thermal generation 229
Thermal generation rate 40
Thermal instability 102, 103
Thermal limitation 489
Thermal noise 117, 118, 271, 398, 666, 670, 673, 688
Thermal property 54
Thermal radiation 601
Thermal resistance 279, 488, 489

Thermal runaway 103, 277, 296
Thermal sensor 3, 743, 744
Thermal velocity 29, 157, 161, 221, 311, 589
Thermally assisted tunneling 718
Thermally sensitive resistor 744
Thermally stable 296
Thermal-relaxation time 709
Thermionic emission 46, 47, 127, 128, 134, 148, 154, 155, 156, 161, 162, 163, 164, 165, 187, 228, 287, 440, 456, 499, 536, 540, 682, 714, 718
Thermionic-emission current 157, 170, 229, 452, 540
Thermionic-emission theory 154, 159, 161, 162, 540
Thermionic-emission-diffusion theory 154, 159, 161
Thermionic-field emission 165, 390, 718
Thermistor 744, 745, 746, 748
Thermocouple 748, 749
Thermoelectric effect 748
Thermoelectric power 54
Thermoelectricity 748
Thermometer 744
THETA 450, 453
  heterojunction 451
Thin-film solar cell 732
Thin-film transistor 343, 345
Thomson effect 748, 749
Thomson EMF 749
Three-dimensional MOSFET 345
Threshold adjust 322
Threshold current 627, 634, 637, 639, 640, 647, 651, 655
Threshold field 514, 515, 521, 531
Threshold gain 627
Threshold voltage 209, 296, 305, 306, 312, 313, 316, 318, 319, 320, 322, 323, 325, 326, 328, 330, 331, 332, 337, 339, 346, 354, 358, 362, 376, 380, 392, 404, 408
Threshold-voltage shift 313
Thyratron 549
Thyristor 2, 123, 548, 549
  bidirectional 577
  field-controlled 591
  gate turn-off 574
  gate-assisted turn-off 576
  light-activated 580
  MIS 447
  static-inductor 591
Time to breakdown 235
TM wave 629
Torque 752
Torsion bar 752
Transconductance (g_m) 254, 262, 275, 280, 296, 306, 309, 312, 337, 365, 384, 385, 390, 391, 392, 395, 396, 406, 410, 541, 548, 585
  extrinsic 254, 348, 395, 407
  intrinsic 254
Transducer 743
Transfer efficiency 707, 709
Transfer inefficiency 707, 708, 709, 710, 711
Transfer resistor 243
Transfer-electron effect 408
Transferred-electron device 2, 510, 511
Transferred-electron effect 2, 37, 382, 386, 396, 510, 514, 515, 524, 534, 536, 537
Transient time 115, 116, 117
Transistor
  analog 586, 590
  bipolar 2, 66, 124, 184, 243, 295, 296, 314, 349, 388, 418, 450, 541, 548, 552, 573, 577, 582, 583, 584, 585, 586, 590, 591, 694, 695, 696, 748, 762
  charge-injection 541
  chemically sensitive field-effect 766
  conductivity-modulated field-effect 582
  depleted-base 590
  DMOS 346, 582, 583
  double-diffused MOS 346
  double-heterojunction bipolar 284, 285
  drift 248
  field-effect 2, 275, 280, 293, 294, 295, 296, 374, 393, 401, 514, 548
  floating-gate tunnel oxide 356
  graded-base bipolar 286
  heterojunction bipolar 282
  heterojunction field-effect 401
  high-electron-mobility 401
  hot-electron 287, 450
  insulated-gate 582
  insulated-gate bipolar 2, 548, 582
  ion-sensitive field-effect 766
  junction 243
  lateral insulated-gate 582

Transistor (*Cont.*)
  laterally diffused MOS 347
  LDMOS 347
  magnetic-field-sensitive field-effect 763
  metal-base 287
  metal-nitride-oxide-silicon 357
  metal-oxide-nitride-oxide-silicon 360
  metal-oxide-semiconductor field-effect 293
  MNOS 357
  modulation-doped field-effect 401
  negative-resistance field-effect 538
  permeable-base 587
  planar-doped-barrier 287
  point-contact 243
  potential-effect 2, 294, 541, 548
  power 279
  real-space-transfer 538
  reset 712
  resonant-tunneling bipolar 459
  resonant-tunneling hot-electron 459
  RST 538
  selectively doped heterojunction 401
  silicon-oxide-nitride-oxide-silicon 360
  single-electron 360
  SONOS 360
  static-induction 586, 591
  thin-film 343, 345
  two-dimensional electron-gas field-effect 401
  velocity-modulation 542
Transistor thermal sensor 748
Transistor-transistor logic 281
Transit angle 475, 480, 493, 502, 503
Transit time 148, 149, 266, 348, 396, 417, 466, 476, 483, 500, 511, 518, 519, 524, 538, 541, 571, 576, 666, 669, 671, 672, 677, 678, 685, 695, 715, 718, 732
Transit-time delay 466, 482
Transit-time device 483
Transit-time dipole-layer mode 525, 526, 527
Transit-time effect 474, 476, 484
Transit-time frequency 486, 492, 525, 526, 527, 529
Transmission coefficient 54, 163, 733
Transmission probability 455
Transport factor 258, 553, 561, 562
Transport velocity 167
Transverse effective mass 425
Transverse electric (TE) wave 628
Transverse field 303, 328
Traveling-wave SOA 657
Triac 577, 578
Triangular barrier 287, 358, 423, 437, 438
Triggering current 571
Triggering temperature 279
Triggering time 278
Triode ac switch 577
Triode-like 591
TTL 281
Tunnel diode 2, 418, 424, 425, 466, 514
  interband 418
  metal-insulator-metal 448
  MIM 448
  MIS 437, 439, 442, 444
Tunneling 2, 47, 102, 103, 104, 107, 153, 161, 163, 166, 227, 287, 330, 336, 351, 353, 354, 356, 357, 359, 417, 607, 608, 692
  band-to-band 120
  coherent 456
  direct 228, 421, 423, 437, 438, 452
  F-N 437, 438
  Fowler-Nordheim 228, 437, 452
  indirect 421, 427
  interband 422
  phonon-assisted 421, 434, 456
  phonon-assisted indirect 427
  resonant 455, 458
  sequential 456
Tunneling current 162, 163, 164, 166, 229, 234, 348, 390, 417, 419, 420, 421, 422, 423, 425, 427, 428, 430, 438, 440, 488, 713
Tunneling hot-electron transfer amplifier 450
Tunneling probability 47, 48, 163, 170, 417, 420, 422, 423, 424, 427, 428, 429, 430, 439, 440, 441, 455
Tunneling resistance 361
Tunneling spectroscopy 434
Tunneling time 417
Tunnel-injection transit-time diode 505
TUNNETT diode 2, 504
Turn-off 594
Turn-off gain 575, 576
Turn-off time 272, 273, 572, 573, 574, 576, 594

Turn-on time 271, 272, 273, 571, 572, 573, 574, 581
Turnover voltage 103
Two-dimensional electron-gas field-effect transistor 401
Two-piece linear approximation 307, 308, 382, 384, 406
Two-sided diode 467

Uncertainty principle 361, 612
Unilateral gain 269, 270, 349, 396, 397
Upper valley 514

Vacuum diode 541
Vacuum tube 549
Valence band 13
Valley current 429, 430, 431, 456, 459
Valley voltage 419, 429, 431
Varactor 120, 121
Variable attenuator 123
Variable reactor 120
Variable resistor 120
Variolosser 123
Varistor 120
VCSEL 655
Velocity modulation 542
Velocity overshoot 37, 310
Velocity saturation 36, 49, 294, 303, 307, 308, 309, 317, 328, 329, 382, 385, 391, 406, 407, 408, 472, 484, 486, 514, 590, 718
Velocity-field relationship 36, 307, 381
Velocity-modulation transistor 542
Vertical transistor 345
Vertical transition 604
Vertical-cavity surface-emitting laser 655
Vibration 753
VMT 542
Voltage regulator 120
Voltage-controlled resistor 375, 376

Wavefunction 12, 47, 48, 58
Waveguide 618, 627
Waveguiding 627, 633
Wavelength converter 613
Wavelength tuning 650
Weak inversion 201, 202, 205, 314
Webster effect 254
Wedge transducer 754
Weight 752
Wentzel-Kramers-Brillouin approximation 48, 422
Wheatstone bridge 751
White noise 118
White-light LED 619
Wigner-Seitz cell 11, 12
WKB approximation 48, 422, 438, 439
Work function 125, 135, 136, 139, 142, 144, 146, 199
Work-function difference 225, 226, 312, 318, 322
Wurtzite lattice 8

Young's modulus 751, 752

Zener diode 120
Zener voltage 120
Zincblende lattice 8, 13